Уральский федеральный университет
имени первого Президента России Б. Н. Ельцина
Институт естественных наук и математики

А. В. ЛОКТИН
А. Б. ОСТРОВСКИЙ

МЕТОДЫ ЗВЕЗДНОЙ СТАТИСТИКИ

Учебное пособие

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
УРАЛЬСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИМЕНИ ПЕРВОГО ПРЕЗИДЕНТА РОССИИ Б. Н. ЕЛЬЦИНА

А. В. Локтин, А. Б. Островский

МЕТОДЫ ЗВЕЗДНОЙ СТАТИСТИКИ

Учебное пособие

Рекомендовано
методическим советом Уральского федерального университета
в качестве учебного пособия для студентов вуза,
обучающихся по направлению подготовки
03.05.01 «Астрономия»

Екатеринбург
Издательство Уральского университета
2018

УДК 521(075.8)
Л 733

Рецензенты:

кафедра астрономии и космической геодезии Национального исследовательского Томского государственного университета (заведующий кафедрой доктор физико-математических наук В. А. Авдюшев);

И. И. Никифоров, кандидат физико-математических наук, доцент кафедры небесной механики Санкт-Петербургского государственного университета

Научный редактор
доктор физико-математических наук, доцент Э. Д. Кузнецов
(Уральский федеральный университет)

Локтин, А. В.

О-777 Методы звездной статистики : учеб. пособие / А. В. Локтин, А. Б. Островский ; [науч. ред. Э. Д. Кузнецов] ; М-во образования и науки Рос. Федерации, Урал. федер. ун-т. — Екатеринбург : Изд-во Урал. ун-та, 2018. — 252 с.

ISBN 978-5-7996-2315-9

В учебном пособии рассмотрены теоретические основы многомерных статистических методов и способы решения практических задач многомерной статистики в применении к проблемам звездной статистики.

Для студентов, обучающихся на астрономических и геодезических специальностях высших учебных заведений.

УДК 521(075.8)

На обложке:
гравюра «Птолемей и муза Астрономия» из книги Reisch Gregor "Margarita Philosophica Cu Additionibus Novis" (1508).

ISBN 978-5-7996-2315-9

ПРЕДИСЛОВИЕ

Основной целью курса «Методы звездной сатистики» является ознакомление студентов с методами, которые используются в звездно-астрономических исследованиях, а также с современными многомерными методами математической статистики, которые находят в астрономии существенно большее применение, чем в недавнем прошлом. Поэтому первые главы посвящены изложению некоторых разделов многомерной математической статистики, которые обычно не включаются в университетские курсы теории вероятностей и математической статистики. В остальных главах рассматриваются звездно-статистические методы, в первую очередь те, которые не излагаются в общем курсе звездной астрономии и в соответствующих учебниках и пособиях.

Изложение материала в рамках курса ведется в векторной и матричной форме, так как большинство современных книг и руководств по математической и звездной статистике используют именно эту наиболее краткую форму. Во всех математических выводах в рамках курса предполагаются непрерывность всех функций, а также существование всех интегралов и производных.

Принятые в курсе обозначения в основном соответствуют сложившейся в математической статистике традиции. Так, случайные величины, векторы и матрицы обозначаются прописными латинскими буквами, значения случайных величин, а также элементы векторов и матриц — строчными латинскими буквами, неслучайные параметры случайных величин — греческими буквами, прописными для векторных и матричных параметров и строчными — для скалярных. В отдельных случаях векторные величины выделяются жирным шрифтом, почти всегда над символами имеется стрелка.

Для обозначения формул принята сквозная нумерация из двух цифр, первая обозначает номер главы, вторая — номер формулы в пределах данной главы.

Объем курса не позволяет детально рассмотреть многие аспекты применения статистических методов в астрономии, поэтому изложение носит часто конспективный характер, отдельные методы изложены без выводов и доказательств. Тем не менее авторы надеются, что настоящий курс окажется полезным и до некоторой степени восполнит отсутствие соответствующего подробного учебника.

Для усвоения курса необходимо знание основ звездной астрономии, по крайней мере, в объеме курса общей астрономии, а также основ теории вероятностей и методов обработки наблюдений. Данное учебное пособие представляет собой переработанное и дополненное пособие А. Е. Василевского [1], на основе которого этот замечательный преподаватель до конца своей жизни читал курс «Методы звездной статистики» на кафедре астрономии и геодезии Уральского государственного университета им. А. М. Горького.

ВВЕДЕНИЕ

Звездная статистика представляет собой раздел астрономии, целью которого является изучение строения, кинематики и эволюции звездных систем — звездных скоплений, нашей Галактики в целом, внегалактических объектов — методами теории вероятностей и математической статистики. В круг задач, решаемых методами звездной статистики, входит также исследование статистических закономерностей и связей между различными физическими и пространственно-кинематическими характеристиками отдельных классов объектов (не только звездных) — лунных и планетных кратеров, образований на поверхности Солнца, комет, астероидов, одиночных и кратных звезд, звездных систем разного масштаба, туманностей, различных классов объектов межзвездной среды, как компактных, так и пространственно распределенных. В последние годы к этим задачам добавилось исследование статистических свойств постоянно растущей выборки экзопланет.

Звездная статистика имеет дело с огромными массивами наблюдательных данных, часто отягощенных большими случайными и систематическими ошибками, поэтому естественно, что этот раздел астрономической науки издавна является областью приложения математической теории обработки резуль-

татов эксперимента. В свою очередь, развитие звездной астрономии в определенной степени влияло и на развитие математической статистики и близких к ней математических дисциплин. Знаменитый метод наименьших квадратов был создан К. Гауссом именно для астрономических приложений. Некоторые чисто статистические приложения, а также широко используемые в прикладной статистике методы решения интегральных уравнений типа свертки были разработаны астрономами Дж. Каптейном, А. Эддингтоном, К. Шварцшильдом и др.

В последние десятилетия наблюдается заметное повышение интереса со стороны астрономов к современным методам математической статистики с целью извлечения максимума информации из массивов наблюдательных данных. Однако и в настоящее время часто встречаются научные работы, где исследования проводятся на недостаточно высоком уровне обработки наблюдательного материала, что ведет к неполному использованию информации, заложенной в наблюдательных данных, получаемых зачастую с большими затратами труда, денежных средств и времени, а иногда и к неправильной интерпретации получаемых результатов. Отсюда следует, что одной из важнейших задач звездной статистики является выделение максимально доступного для данного массива наблюдательных данных количества полезной информации. С этим напрямую связан вопрос о планировании эксперимента в звездной астрономии, хотя эксперимент в звездной астрономии является пассивным, так как мы не способны непосредственно воздействовать на изучаемые объекты. Однако выяснение вопроса, какое количество данных необходимо для изучения того или иного явления, а также точность, с которой должны получаться эти

данные, может сэкономить при достижении нужного нам результата много усилий.

Звездная астрономия XX в., за исключением его последних десятилетий, имела дело почти исключительно с одномерными случайными величинами и одномерными распределениями. Но по мере развития средств вычислительной техники многомерный статистический анализ превратился из теоретического раздела математической статистики в мощный инструмент научных исследований, в средство извлечения максимальной информации из экспериментальных данных. Многомерные статистические методы проникли во все области знания, в том числе и в традиционно гуманитарные науки — социологию, психологию, лингвистику, широко применяются в экономике. К сожалению, в астрономию эти методы внедряются медленно, что связано, как справедливо отмечено А. С. Шаровым в предисловии к книге Р. Курта «Введение в звездную статистику» [2], со слабой подготовкой астрономов в области теории вероятностей и математической статистики. Данный курс в некоторой мере преследует цель восполнить, хотя бы частично, этот пробел.

Следует сказать несколько слов и о современных вычислительных возможностях в области математической статистики. Особо необходимо отметить язык высокого уровня и систему статистических вычислений R, развитие которых поддерживается сообществом Comprehensive R Archive Network (CRAN). Важно, что соответствующие пакеты программ реализованы для всех основных операционных систем (Windows, Linux, MacOS). Описание языка и принципов работы с ним распространяются бесплатно и постоянно обновляются. В настоя-

щее время, кроме базовых пакетов, скачиваемых вместе с системой, разработаны и доступны более 2 000 прикладных пакетов, реализующих практически все самые современные статистические методы для применения в самых разных областях науки. Скачать программу и пакеты можно с сайта проекта CRAN http://cran.r-project.org и многочисленных сайтов, посвященных R, которые легко найти во Всемирной паутине.

1. МНОГОМЕРНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

1.1. Случайный вектор и его распределение

Определение 1.1. *Случайные величины могут быть скалярными (одномерными) и векторными. В соответствии с общим определением* ***многомерной случайной величиной*** *или* ***случайным вектором*** *называют любую упорядоченную совокупность скалярных случайных величин, причем эти составляющие могут быть как независимыми, так и коррелированными — связанными некоторыми, часто скрытыми от нас, зависимостями.*

В практических приложениях случайные векторы представляют собой, как правило, упорядоченные наборы характеристик или признаков случайно выбранного объекта рассматриваемого класса. В качестве примера мы можем рассмотреть основные характеристики звезды (масса, радиус, светимость, эффективная температура, возраст), причем хорошо известно, что эти величины в разной степени коррелированы. Иногда многомерными случайными величинами являются реаль-

ные векторы, например, вектор положения объекта в Галактике, вектор пространственной скорости объекта.

Любой вектор может быть записан в виде вектора-столбца, или, что удобнее для записи, вектора-строки (транспонированного вектора-столбца), например, $\vec{\mathbf{X}}^T = (x_1,\ x_2,\ \ldots,\ x_m)$, где верхний индекс T обозначает операцию транспонирования. Вектор-столбец удобнее для записи умножения вектора на матрицу, так как запись получается более компактной. Одномерные составляющие вектора называются его компонентами, проекциями или координатами. Геометрически компоненты случайного вектора задают положение точки в m-мерном пространстве, причем, в силу случайного характера величин $x_1,\ x_2,\ \ldots,\ x_m$, положение этой точки от опыта к опыту меняется непредсказуемым образом в соответствии с законом распределения вектора $\vec{\mathbf{X}}$.

Определение 1.2. ***Законом распределения*** *многомерной случайной величины $\vec{\mathbf{X}}$ называется всякое соотношение, устанавливающее связь между возможными значениями этой случайной величины и соответствующими им вероятностями. Наиболее общей формой закона распределения является* ***функция распределения*** *вероятности случайной величины.*

Определение 1.3. ***Функция распределения*** *F случайного вектора $\vec{\mathbf{X}}$ (совместная функция распределения) есть вероятность попадания точки в m-мерную, неограниченную слева область с вершиной в точке $\vec{\boldsymbol{\alpha}}$, причем аргументами функции являются координаты вершины $\vec{\boldsymbol{\alpha}}(\alpha_1,\ \ldots,\ \alpha_m)$ этой области:*

$$F(\alpha_1,\ \ldots,\ \alpha_m) = P(x_1 < \alpha_1,\ \ldots,\ x_m < \alpha_m), \qquad (1.1)$$

где $(\alpha_1,\ \ldots,\ \alpha_m)$ *— (неслучайный!) вектор аргументов функции распределения;* P *— вероятность попадания в область.*

Из последнего выражения ясно, что область значений функции распределения — это интервал от 0 до 1, а областью определения по каждой из переменных может быть как вся числовая ось (это предполагается ниже при записи всех интегралов), так и некоторый интервал числовой оси. При этом F является *неубывающей функцией* своих аргументов во всей области определения. Вероятность для события $\vec{\mathbf{X}}$ реализоваться в пересечении областей, задаваемых некоторыми значениями $\vec{\boldsymbol{\alpha}}$ и $\vec{\boldsymbol{\beta}}$, определяется через значения $F(\vec{\boldsymbol{\alpha}})$, $F(\vec{\boldsymbol{\beta}})$ и равна приращению функции распределения на соответствующем пересечении областей:

$$P\left(\vec{\mathbf{X}},\ \vec{\boldsymbol{\alpha}},\ \vec{\boldsymbol{\beta}}\right) = \left|F(\vec{\boldsymbol{\alpha}}) - F(\vec{\boldsymbol{\beta}})\right|. \tag{1.2}$$

Определение 1.4. ***Плотность вероятности*** $f(x_1,\ \ldots,\ x_m)$ *распределения многомерного вектора может быть найдена дифференцированием функции распределения по ее аргументам:*

$$f(x_1,\ \ldots,\ x_m) = \frac{\partial^m F(x_1,\ \ldots,\ x_m)}{\partial x_1\ \ldots\ \partial x_m}. \tag{1.3}$$

Для получения функции распределения по его плотности следует провести обратное преобразование:

$$F(\alpha_1,\ \ldots,\ \alpha_m) = \int\limits_{-\infty}^{\alpha_1} \cdots \int\limits_{-\infty}^{\alpha_m} f(t_1,\ \ldots,\ t_m)\, dt_1 \ldots dt_m. \tag{1.4}$$

Многомерная плотность вероятности характеризует вероятность попадания точки (конца многомерного

радиус-вектора) в m-мерный параллелепипед со сторонами $(x_1, \ x_1 + dx_1), \ \ldots, \ (x_m, \ x_m + dx_m)$. Вероятность попадания в дифференциально малый параллелепипед равна $f(x_1, \ \ldots, \ x_m)\, dx_1 \ldots dx_m$. Таким образом, для бесконечно малого (дифференциального) многомерного объема плотность вероятности вводится как коэффициент пропорциональности между величиной объема и вероятностью для случайной величины попасть в него.

Бесконечные пределы в интегралах выражения (1.4) и других (см. ниже) появляются на практике не всегда и определяются областью определения аргументов. Так, функции светимости как плотности распределения светимостей звезд (чаще — абсолютных звездных величин) сосредоточены на интервале, определяемом пределами наблюдаемых светимостей звезд, приблизительно от -11 до $+17^m$. Массы наблюдаемых в нашей Галактике звезд, определяющие звездные величины, сосредоточены в интервале от, приблизительно, 60 солнечных масс до, также приблизительно, 0.08 солнечных масс. Остаточные скорости звезд в Галактике не превышают 400 км/с и т. д.

Примером многомерной плотности вероятности может служить эллипсоидальный **закон Шварцшильда** [3] для распределения остаточных скоростей звезд, который, по своей сути, является трехмерным нормальным распределением:

$$f(\vec{\mathbf{V}}) = \frac{1}{\sqrt{(2\pi)^3 \left\| \hat{\Sigma}_{\vec{\mathbf{V}}} \right\|}} \cdot \exp\left(-\frac{1}{2}\, \vec{\mathbf{V}}^T\, \hat{\Sigma}^{-1}_{\vec{\mathbf{V}}}\, \vec{\mathbf{V}} \right), \qquad (1.5)$$

где $\vec{\mathbf{V}}^T = (u, \ v, \ w)$ — вектор пространственной скорости звез-

ды; $\hat{\Sigma}_{\vec{\mathbf{V}}}$ есть так называемая **ковариационная матрица** вектора $\vec{\mathbf{V}}$; $\left\|\hat{\Sigma}_{\vec{\mathbf{V}}}\right\|$ — определитель этой матрицы.

Отметим, что для трехмерного нормального распределения поверхности равной плотности вероятности есть эллипсоиды.

Ковариационная матрица вектора $\vec{\mathbf{V}}$ определяется как

$$\begin{aligned}\hat{\Sigma}_{\vec{\mathbf{V}}} &= \operatorname{cov}\left(\vec{\mathbf{V}},\ \vec{\mathbf{V}}\right) \equiv \operatorname{cov}\left(\vec{\mathbf{V}}\right) = \\ &= \hat{\mathrm{E}}\left(\left(\vec{\mathbf{V}} - \hat{\mathrm{E}}\vec{\mathbf{V}}\right)\left(\vec{\mathbf{V}} - \hat{\mathrm{E}}\vec{\mathbf{V}}\right)^{T}\right) = |\sigma_{ij}|\,,\end{aligned} \tag{1.6}$$

$$\begin{aligned}\sigma_{ij} &= \operatorname{cov}\left(v_i,\ v_j\right) = \\ &= \hat{\mathrm{E}}\left(\left(v_i - \hat{\mathrm{E}}v_i\right)\left(v_j - \hat{\mathrm{E}}v_j\right)\right), \quad i,\ j = 1,\ \ldots,\ n,\end{aligned} \tag{1.7}$$

где $\hat{\mathrm{E}}$ — оператор математического ожидания.

Такая матрица ковариации является обобщением **дисперсии** для многомерной случайной величины, а ее след — скалярным выражением дисперсии многомерной случайной величины. Собственные векторы и собственные числа этой матрицы позволяют оценить размеры и форму облака распределения такой случайной величины, аппроксимировав его эллипсоидом (или эллипсом в двумерном случае, как показано на рис. 1.1 и 1.2).

На практике часто возникает необходимость найти **функцию распределения проекции случайного вектора** на одну из координатных осей, например, на ось x_1. По определению искомая функция есть

$$F_1(\alpha_1) = P(x_1 < \alpha_1) \tag{1.8}$$

для всего диапазона значений прочих параметров. Нижний индекс при функции распределения обозначает размерность про-

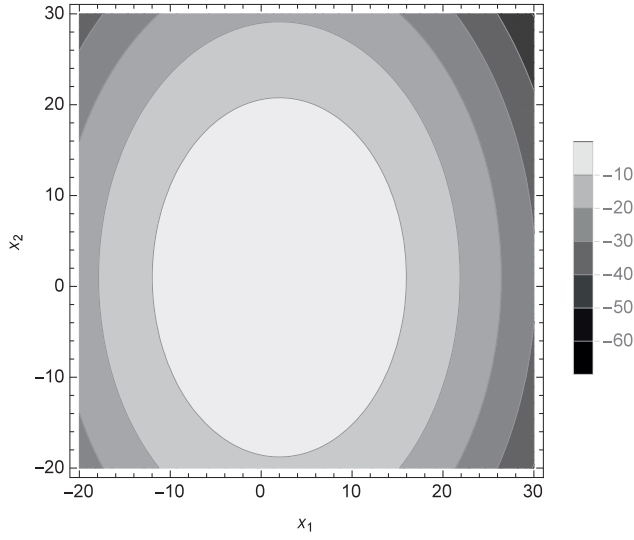

Рис. 1.1. Натуральный логарифм плотности вероятности двумерного нормального распределения (серая шкала). Математическое ожидание задается вектором (2, 1). Дисперсия по горизонтальной оси $\sigma_{11} = 1$, по вертикальной $\sigma_{22} = 2$. Случайные величины некоррелированы

екции (1 — проецирование на координатную ось, то есть пространство единичной размерности).

Соответственно для плотности распределения проекции случайного вектора имеем следующее определение.

Определение 1.5. ***Плотность распределения проекции случайного вектора*** *равна*

$$f_1(\alpha_1) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(\alpha_1,\, x_2,\, \ldots,\, x_m)\, dx_2 \ldots dx_m. \qquad (1.9)$$

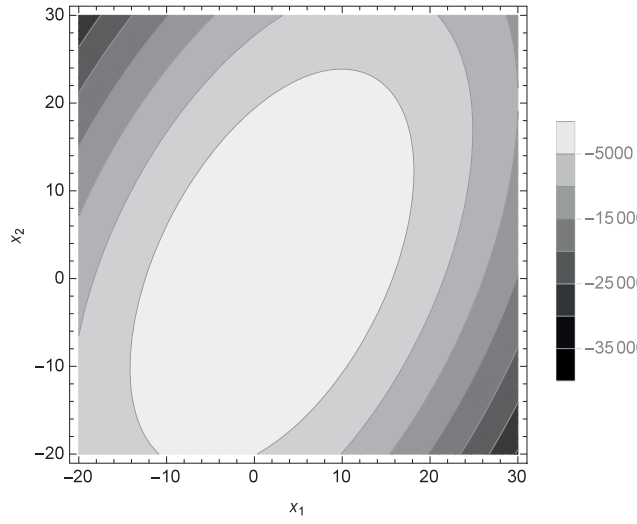

Рис. 1.2. Натуральный логарифм плотности вероятности двумерного нормального распределения (серая шкала). Математическое ожидание задается вектором (2, 1). Дисперсия по горизонтальной оси $\sigma_{11} = 1$, по вертикальной $\sigma_{22} = 2$. Коэффициент корреляции случайных величин $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = -0.99$

Бесконечные пределы интегрирования в (1.9) *указаны условно и подразумевают всю область определения соответствующих параметров.*

Аналогично можно найти плотность распределения проекции случайного вектора $\vec{\mathbf{X}}$ на любую из осей и вообще на любое k-мерное подпространство ($k < m$). Плотность распределения проекции в этом случае будет обозначаться как $f_k(\alpha_1, \ldots, \alpha_k)$. Такие распределения называют **частными** или **маргинальными**.

Отметим, что для случайного вектора со статистически независимыми (некоррелированными) компонентами многомерная плотность вероятности равна произведению его частных одномерных плотностей.

Примерами частных распределений в звездной астрономии могут служить видимое распределение звездной плотности в скоплениях (проекция их пространственной плотности на картинную плоскость), распределение лучевых скоростей звезд в малой площадке неба (проекция распределения пространственных скоростей звезд на луч зрения).

Если зафиксировать значения некоторых координат случайного вектора, например $x_{k+1} = \alpha_{k+1},\ \ldots,\ x_m = \alpha_m$, то получим уже не проекцию, а **сечение** m-мерного распределения k-мерной гиперплоскостью. Плотность распределения оставшихся незафиксированными координат называется k-мерной **условной плотностью распределения**. На основе теоремы умножения вероятностей легко показать, что нормированная условная плотность определяется следующим образом.

Определение 1.6. ***Нормированная условная k-мерная плотность распределения*** *равна*

$$f_{(k)} = f_{(k)}\left(x_1,\ \ldots,\ x_k \,|\, \alpha_{k+1},\ \ldots,\ \alpha_m\right) = \\ = \frac{f\left(x_1,\ \ldots,\ x_k,\ \alpha_{k+1},\ \ldots,\ \alpha_m\right)}{f_{m-k}\left(\alpha_{k+1},\ \ldots,\ \alpha_m\right)}, \tag{1.10}$$

где $x_1,\ \ldots,\ x_k$ — независимые аргументы; $\alpha_{k+1},\ \ldots,\ \alpha_m$ — зафиксированные значения аргументов, то есть определенные числа.

Примерами условных распределений являются функция светимости звезд фиксированного возраста, распределение звездной плотности в направлении полюса Галактики (сечение пространственного распределения звезд прямой линией).

Некоторые многомерные распределения, в частности нормальное распределение, при преобразованиях проекции или сечения не изменяют своего вида. Например, если распределение пространственных скоростей звезд является нормальным, то и распределения тангенциальных и лучевых скоростей тоже будут нормальными.

Определение 1.7. *Случайные величины* $x_1,\ x_2,\ \ldots,\ x_n$ *называются* ***статистически независимыми****, если функция их совместного распределения* $F(x_1,\ x_2,\ \ldots,\ x_n) = F(x_1)F(x_2)\ldots F(x_n)$.

1.2. Моменты случайного вектора

Определение 1.8. *Начальным моментом первого порядка или* ***математическим ожиданием*** *случайной многомерной величины* $\vec{\mathbf{X}}$ *называется вектор (неслучайный)* $\vec{\boldsymbol{\xi}} = (\xi_1,\ \ldots,\ \xi_m)$, *компонентами которого являются математические ожидания компонент вектора* $\vec{\mathbf{X}}$:

$$\begin{aligned} \xi_j = \hat{\mathrm{E}}(x_j) &= \int\limits_{-\infty}^{+\infty} x_j f_1(x_j)\,dx_j = \\ &= \int\limits_{-\infty}^{+\infty} \cdots \int\limits_{-\infty}^{+\infty} x_j f(x_1,\ \ldots,\ x_m)\,dx_1\ \ldots\ dx_m, \end{aligned} \tag{1.11}$$

где интегрирование ведется по всей области определения плотности вероятности f, а $\hat{\mathrm{E}}$ — оператор математического ожидания.

Полезно помнить, что $\hat{\mathrm{E}}$ является линейным оператором и, с учетом определения математического ожидания (1.11), для него выполняются следующие равенства:

$$\hat{\mathrm{E}}\left(\alpha\vec{\mathbf{X}}\right) = \alpha\,\hat{\mathrm{E}}\left(\vec{\mathbf{X}}\right), \tag{1.12}$$

$$\hat{\mathrm{E}}\left(\vec{\mathbf{X}} + \vec{\mathbf{Y}}\right) = \hat{\mathrm{E}}\left(\vec{\mathbf{X}}\right) + \hat{\mathrm{E}}\left(\vec{\mathbf{Y}}\right), \tag{1.13}$$

где α — число.

Определение 1.9. *Случайную величину называют* ***центрированной****, если ее математическое ожидание равно нулю.*

Определение 1.10. *Центральным моментом второго порядка для m-мерного случайного вектора является $(m \times m)$-матрица вторых центральных моментов всех компонент вектора, называемая* ***ковариационной матрицей****:*

$$\hat{\Sigma}_{\vec{\mathbf{X}}} = \hat{\mathrm{E}}\left(\left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}}\right)\left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}}\right)^T\right) = \|\sigma_{ij}\|, \tag{1.14}$$

причем будем обозначать

$$\sigma_{jk} = \hat{\mathrm{E}}\left(\left(x_j - \xi_j\right)\left(x_k - \xi_k\right)\right), \tag{1.15}$$

*где σ_{jj} — **дисперсии**; σ_{jk} при $j \neq k$ называются **ковариациями** элементов вектора $\vec{\mathbf{X}}$. Иначе $\sigma_{jj} \equiv \sigma_j^2$, где σ_j называют **стандартным отклонением**; $\sigma_{jk} \equiv \mathrm{cov}(x_j,\, x_k)$.*

Определение 1.11. ***Корреляционная матрица*** $\hat{\mathrm{P}}_{\vec{\mathbf{X}}}$ *может быть получена из ковариационной матрицы, для чего элементы матрицы* $\hat{\Sigma}_{\vec{\mathbf{X}}}$ *необходимо заменить на*

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\,\sigma_{kk}}} = \frac{\operatorname{cov}(x_j,\, x_k)}{\sqrt{\sigma_j^2\,\sigma_k^2}}; \quad \rho_{jj} \equiv 1. \tag{1.16}$$

Ковариационная матрица является обобщением понятия **дисперсии** на случай многомерного распределения и является *мерой рассеяния точек относительно математического ожидания в многомерном пространстве*. Иногда *определитель ковариационной матрицы* называют **обобщенной дисперсией**.

Из определения матриц $\hat{\Sigma}_{\vec{\mathbf{X}}}$ и $\hat{\mathrm{P}}_{\vec{\mathbf{X}}}$ видно, что они *квадратные*, *симметричные* и имеют *неотрицательные элементы на главной диагонали*. Кроме того, они являются *неотрицательно определенными*.

Для случайного вектора с некоррелированными компонентами обе матрицы диагональны (ковариации равны нулю), при этом корреляционная матрица превращается в единичную матрицу.

Ранг ковариационной матрицы определяет степень вырожденности распределения. Например, если у трехмерного случайного вектора ранг матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$ равен двум, то распределение вырождено до двумерного, то есть вся масса вероятности лежит в одной плоскости.

В силу неотрицательной определенности матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$ уравнение вида

$$\left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}}\right)^T \hat{\Sigma}_{\vec{\mathbf{X}}}^{-1} \left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}}\right) = \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{\Sigma_{jk}}{\left\|\hat{\Sigma}_{\vec{\mathbf{X}}}\right\|} x_j \, x_k = m + 2\,, \quad (1.17)$$

где Σ_{jk} — алгебраическое дополнение элемента σ_{jk} (минор, полученный вычеркиванием j-й строки и k-го столбца, домноженный на $(-1)^{j+k}$), представляет собой неотрицательно определенную квадратичную форму, геометрическим образом которой является m-мерный гиперэллипсоид, называемый **эллипсоидом рассеяния** [4]. Квадратичная форма, соответствующая (1.17), задает область, ограниченную m-мерным гиперэллипсоидом с постоянной плотностью вероятности. Заданный таким образом эллипсоид рассеяния соответствует исходному распределению вероятностей вектора $\vec{\mathbf{X}}$, в том смысле, что равномерное распределение случайной величины в гиперобъеме этого эллипсоида имеет моменты первого и второго порядков, равные соответствующим моментам вектора $\vec{\mathbf{X}}$. Эллипсоид рассеяния дает геометрическое описание расположения массы плотности вероятности в пространстве. Его размеры в направлениях осей равны $2\,\sigma_j$, а ориентация определяется внедиагональными элементами матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$. Если эта матрица диагональна, то главные оси эллипсоида рассеяния параллельны осям координат, а по длине равны удвоенным значениям соответствующих дисперсий.

1.3. Условное математическое ожидание (регрессия)

Определение 1.12. ***Условным математическим ожиданием*** *одной из компонент случайного вектора* $\vec{\mathbf{X}}$ *(в данном случае* m*-й) называется неслучайная величина*

$$\xi_m(x_1, \ldots, x_{m-1}) = \hat{\mathrm{E}}(x_m \mid x_1, \ldots, x_{m-1}) = \\ = \int_{-\infty}^{+\infty} x_m f_{(1)}(x_m \mid x_1, \ldots, x_{m-1})\, dx_m. \tag{1.18}$$

Из выражения (1.18) видно, что условное математическое ожидание зависит от конкретных значений переменных $x_1, \ldots, x_{m-1}$, то есть является функцией этих переменных. В теории вероятностей такую функцию называют **регрессией** x_m на $x_1, \ldots, x_{m-1}$. Для m-мерного случайного вектора можно построить m регрессий вида (1.18), то есть для каждой из переменных в отдельности. Каждое уравнение регрессии описывает некоторую поверхность, называемую поверхностью регрессии. Для многомерного нормального распределения поверхности регрессии являются гиперплоскостями, а для двумерного нормального распределения — прямыми линиями. Последний случай иллюстрируется на рис. 1.3—1.5, на которых изображены эллипс рассеяния и все возможные регрессии, которых в случае двумерного распределения будет две.

Особое место в математической статистике занимает так называемая **линейная среднеквадратическая регрессия**.

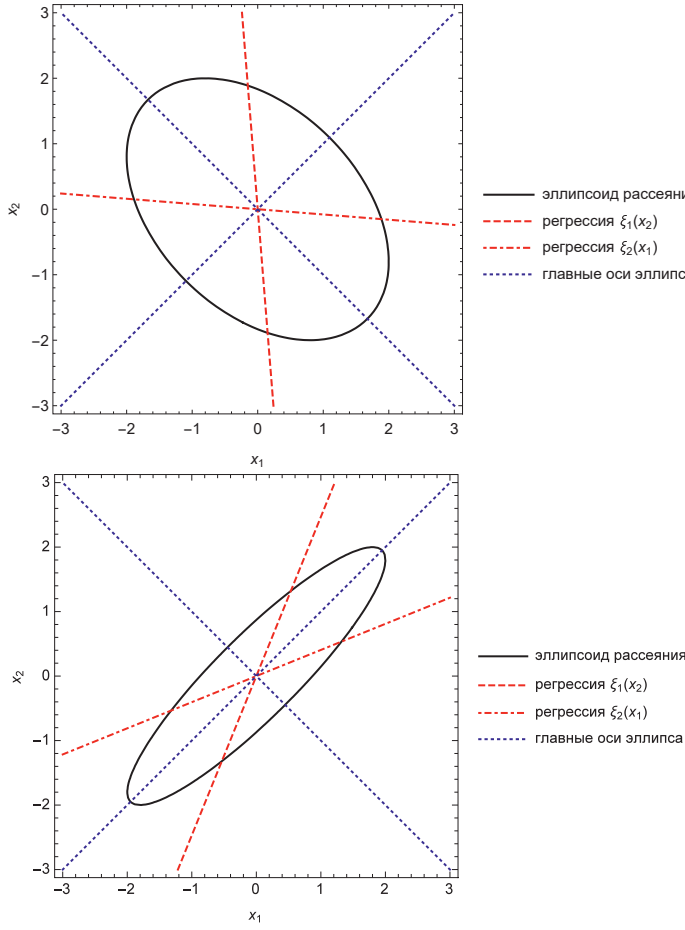

Рис. 1.3. Эллипс рассеяния и все возможные регрессии для двумерного нормального распределения. Матрица $\hat{\Sigma}_{\vec{\mathbf{X}}}$ задана своими элементами $\sigma_{11} = 1$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = -0.4$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = -0.4$ (верхняя панель); $\sigma_{11} = 1$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 0.9$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = 0.9$ (нижняя панель)

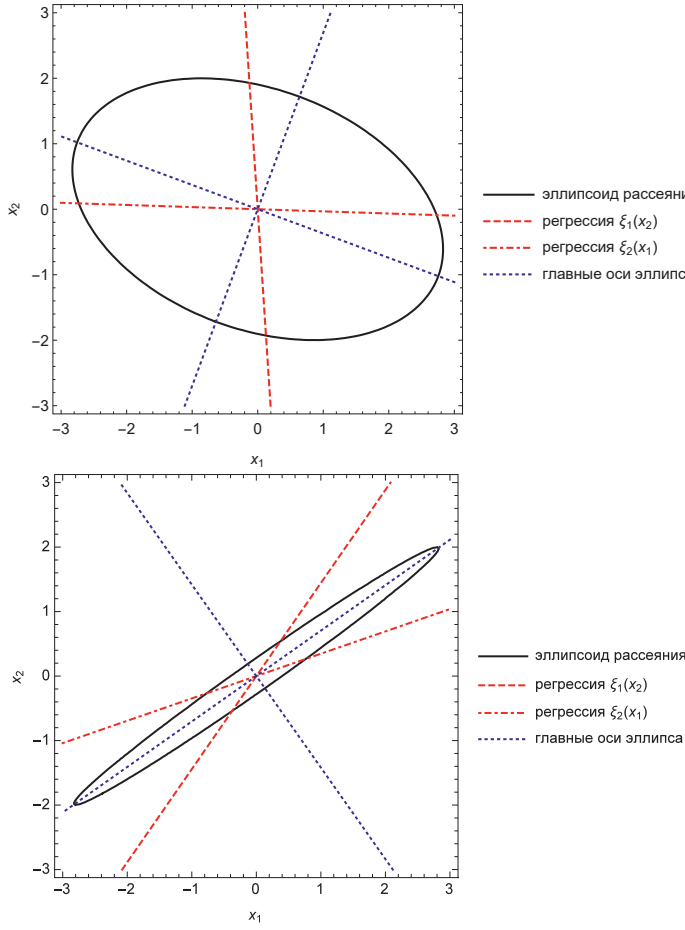

Рис. 1.4. Эллипс рассеяния и все возможные регрессии для двумерного нормального распределения. Матрица $\hat{\Sigma}_{\vec{\mathbf{X}}}$ задана своими элементами $\sigma_{11} = 2$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = -0.43$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = -0.3$ (верхняя панель); $\sigma_{11} = 2$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 1.4$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = 0.99$ (нижняя панель)

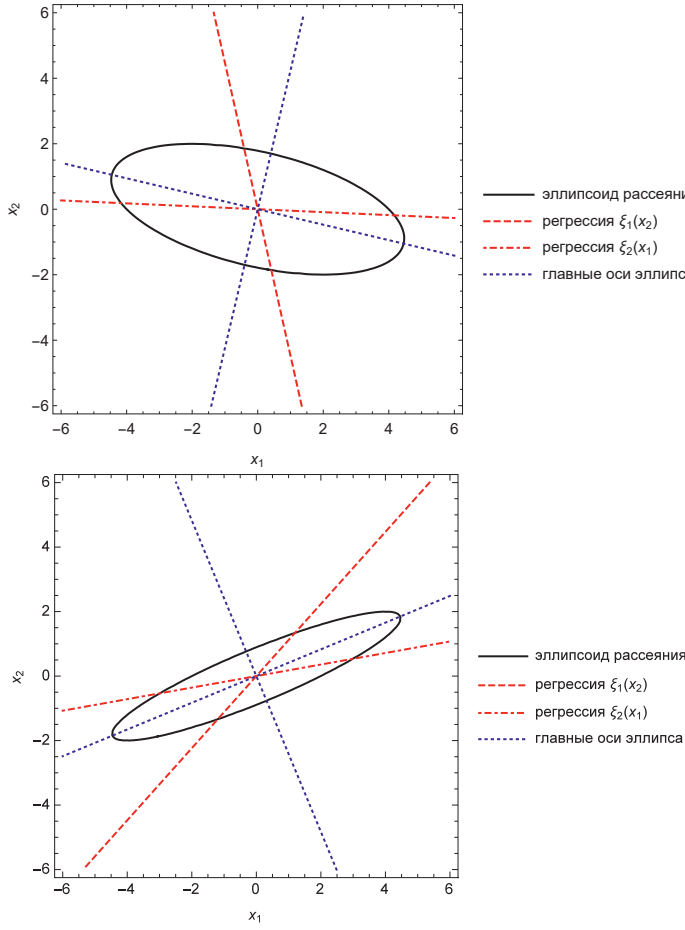

Рис. 1.5. Эллипс рассеяния и все возможные регрессии для двумерного нормального распределения. Матрица $\hat{\Sigma}_{\vec{\mathbf{x}}}$ задана своими элементами $\sigma_{11} = 5$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = -1.0$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = -0.45$ (верхняя панель); $\sigma_{11} = 5$, $\sigma_{22} = 1$, $\sigma_{12} = \sigma_{21} = 2.0$, коэффициент корреляции $\sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = 0.9$ (нижняя панель)

Определение 1.13. *Для координаты x_m случайного вектора $\vec{\mathbf{X}}$ уравнение* ***линейной среднеквадратической регрессии*** *имеет вид*

$$\xi_m^0 = \beta_{m,\,1}(x_1 - \xi_1) + \ldots + \beta_{m,\,m-1}(x_{m-1} - \xi_{m-1})\,, \qquad (1.19)$$

где $\beta_{m,\,k}$ — постоянные коэффициенты, которые в данном случае определяются по способу наименьших квадратов, то есть из условия

$$\hat{\mathrm{E}}\left(x_m - (\beta_{m,\,1}(x_1 - \xi_1) + \ldots + \beta_{m,\,m-1}(x_{m-1} - \xi_{m-1}))\right)^2 = \\ = \hat{\mathrm{E}}\left(x_m - \xi_m^0\right)^2 = \min\,. \qquad (1.20)$$

Можно показать [5], что для *невырожденного распределения* коэффициенты

$$\beta_{jk} = \frac{\Sigma_{jk}}{\Sigma_{jj}}\,, \qquad (1.21)$$

где Σ_{jk} есть алгебраическое дополнение jk-го элемента определителя ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$. Если вектор $\vec{\mathbf{X}}$ имеет некоррелированные компоненты, то все внедиагональные алгебраические дополнения и все коэффициенты β_{jk} $(j \neq k)$ равны нулю.

Уравнению (1.19) соответствует гиперплоскость среднеквадратической регрессии, которая дает *наилучшую линейную аппроксимацию* распределения, то есть величина ξ_m^0 является наилучшей в смысле минимума выражения (1.20) линейной оценкой случайной величины x_m по величинам $x_1,\ \ldots,\ x_{m-1}$, *рассматриваемым как независимые переменные.*

1.4. Парные, частные и множественный коэффициенты корреляции

Определение 1.14. ***Парный коэффициент корреляции (коэффициент корреляции Пирсона; r Пирсона; простая корреляция)*** *является мерой близости связи двух координат x_j и x_k случайного вектора $\vec{\mathbf{X}}$ к линейной функциональной зависимости. При этом возможное влияние остальных компонент вектора $\vec{\mathbf{X}}$ не принимается во внимание и не исключается (то есть мерой близости связи к линейной зависимости при неисключенном влиянии на эту связь других координат вектора, а это влияние может быть синхронным на координаты x_j и x_k (коррелированным)).*

Парный коэффициент корреляции рассчитывается следующим образом [6]:

$$r_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\,\sigma_{kk}}} \approx \approx \frac{\sum\limits_{i=1}^{n} \left(x_j^i - \langle x_j \rangle\right)\left(x_k^i - \langle x_k \rangle\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(x_j^i - \langle x_j \rangle\right)^2 \sum\limits_{i=1}^{n} \left(x_k^i - \langle x_k \rangle\right)^2}}, \tag{1.22}$$

где n — число измерений вектора $\vec{\mathbf{X}}$ в серии наблюдений; индекс i обозначает номер измерений в серии; $\langle x_l \rangle$ — среднее по серии измерений значение координаты вектора $\vec{\mathbf{X}}$.

Определение 1.15. *Назовем* ***частным коэффициентом корреляции*** *показатель близости связи двух координат случайного вектора* $\vec{\mathbf{X}}$ *к линейной функциональной зависимости* после исключения влияния остальных его координат.

По сути, частный коэффициент корреляции может быть построен как обычный парный коэффициент корреляции, но ковариационная матрица должна соответствовать условной функции распределения (например, для компонент x_1 и x_2) $f_{(2)}(x_1,\ x_2 \,|\, \alpha_3,\ \ldots,\ \alpha_m)$, определенной согласно (1.10). Именно в таком смысле учитывается влияние компонент $x_3,\ \ldots,\ x_m$ вектора $\vec{\mathbf{X}}$ на компоненты x_1 и x_2.

Вычисление частного коэффициента корреляции $\rho_{1,\,2\,|\,3,\,4,\,\ldots,\,m}$ для двух компонент x_1 и x_2 производится следующим образом.

1. По формуле (1.19) вычисляют линейные регрессии $\xi_1^0(x_3,\ x_4,\ \ldots,\ x_m)$ и $\xi_2^0(x_3,\ x_4,\ \ldots,\ x_m)$, в которых заключено влияние остальных компонент вектора $\vec{\mathbf{X}}$ на компоненты x_1 и x_2. При использовании формулы (1.19) вместо математических ожиданий соответствующих величин следует использовать их выборочные средние значения.
2. Вычисляют свободные от влияния переменных $x_3,\ x_4,\ \ldots,\ x_m$ разности $z_1 = x_1 - \xi_1^0$ и $z_2 = x_2 - \xi_2^0$, которые также являются случайными величинами.
3. Вычисляют парный коэффициент корреляции между z_1 и z_2, который и будет частным коэффициентом корреляции $\rho_{1,\,2\,|\,3,\,4,\,\ldots,\,m}$ для x_1 и x_2.

Более практичным является следующий способ вычисления частных коэффициентов корреляции. Можно показать, что

для любых двух координат x_j и x_k случайного вектора $\vec{\mathbf{X}}$

$$\rho_{jk} = -\frac{P_{jk}}{\sqrt{P_{jj}P_{kk}}} = -\frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}\,, \tag{1.23}$$

где Σ_{jk} и P_{jk} — алгебраические дополнения соответствующих элементов ковариационной и корреляционной матриц $\hat{\Sigma}_{\vec{\mathbf{X}}}$, $\hat{\mathrm{P}}_{\vec{\mathbf{X}}}$. Из величин ρ_{jk} можно составить матрицу частных коэффициентов корреляции.

Если парный коэффициент корреляции между данными случайными величинами отличен от соответствующего частного коэффициента, то, следовательно, фиксированные величины усиливают (или ослабляют) взаимосвязь между изучаемыми переменными.

Определение 1.16. ***Множественный коэффициент корреляции*** *ρ_j выполняет ту же функцию, что и парный, то есть служит мерой линейной связи, но между одним из элементов случайного вектора и всеми остальными вместе. Он определяется как парный коэффициент корреляции между x_j и ξ_j^0 и обычно вычисляется по формуле*

$$\rho_j^2 = 1 - \frac{\left\|\hat{\mathrm{P}}\right\|}{P_{jj}} = 1 - \frac{\left\|\hat{\Sigma}_{\vec{\mathbf{X}}}\right\|}{\sigma_j^2\,\Sigma_{jj}}\,. \tag{1.24}$$

В приложениях часто используют именно квадрат множественного коэффициента корреляции, иногда называемый **коэффициентом детерминации**. Желательно запомнить, что если ρ_j^2 равен единице, то случайная величина x_j является некоторой линейной комбинацией остальных элементов случайного вектора $\vec{\mathbf{X}}$, то есть распределение вырождено по меньшей мере

в отношении координаты x_j. Если $\rho_j^2 = 0$, то этот компонент случайного вектора не коррелирован ни с одним из оставшихся компонентов. Если же все $\rho_j^2 = 0$ равны нулю, то корреляционная матрица вектора превращается в единичную.

При анализе коррелированности случайных величин должен рассматриваться вопрос о значимости полученного коэффициента корреляции. Необходимо проверить статистическую гипотезу $H_0 : \rho_{ij} = 0$. Если эта гипотеза будет отвергнута, то это означает наличие линейной зависимости между случайными величинами. В качестве статистики, с помощью которой проверяется гипотеза H_0, может быть использована следующая функция [7]

$$T = \frac{\rho_{ij}\sqrt{n-2}}{\sqrt{1-\rho_{ij}^2}}, \tag{1.25}$$

где ρ_{ij} — коэффициент корреляции, вместо него может быть использован выборочный коэффициент корреляции (2.7). Функция T удовлетворяет распределению Стьюдента с $(n-2)$ степенями свободы. Для заданного уровня значимости α гипотеза H_0 отвергается, если $|\,T\,| \geqslant t_\alpha$, где t_α — α-процентная точка распределения Стьюдента с $(n-2)$ степенями свободы.

По своему смыслу квадрат множественного коэффициента корреляции равен доле дисперсии данного компонента случайного вектора, которая «объясняется» регрессионной зависимостью с остальными компонентами. Последнее определяет применение множественного коэффициента корреляции в дисперсионном и регрессионном анализе, что будет ясно из дальнейшего. По определению множественный коэффициент корреляции неотрицателен.

1.5. Функции случайного вектора

В практических приложениях бывает необходимо найти распределение случайного вектора, функционально связанного с другим случайным вектором, распределение которого известно. Много примеров решения этой задачи встретятся в последующих главах. Очевидным примером является переход от функции светимости к функции масс с помощью зависимости масса—светимость.

Пусть случайный вектор $\vec{\mathbf{Y}}^T = (y_1, \ldots, y_n)$ является функцией от случайного вектора $\vec{\mathbf{X}}^T = (x_1, \ldots, x_m)$. Это означает, что каждый элемент вектора $\vec{\mathbf{Y}}$ функционально связан со всеми элементами вектора $\vec{\mathbf{X}}$:

$$y_j = \varphi_j(x_1, \ldots, x_m), \quad j = 1, \ldots, n. \tag{1.26}$$

В общем случае $n \neq m$, но для простоты рассмотрим частный случай $n = m$. Точные формулы такого преобразования сложны, поэтому рассмотрим применяемые в практических приложениях приближенные формулы, следующие из разложения функций $\varphi(\vec{\mathbf{X}})$ в ряд в окрестностях точки $\vec{\boldsymbol{\xi}}$, и тогда

$$\hat{\mathrm{E}}\left(\vec{\mathbf{Y}}\right) \equiv \vec{\boldsymbol{\eta}} \approx \varphi\left(\vec{\boldsymbol{\xi}}\right), \tag{1.27}$$

$$\hat{\Sigma}_{\vec{\mathbf{Y}}} \approx \hat{\mathrm{D}}\, \hat{\Sigma}_{\vec{\mathbf{X}}}\, \hat{\mathrm{D}}^T, \tag{1.28}$$

где $\hat{\mathrm{D}}$ — $(m \times m)$-матрица частных производных в точке $\vec{\boldsymbol{\xi}}$ от функции $\varphi\left(\vec{\mathbf{X}}\right)$ по всем x_j. Точность формул (1.28) тем выше, чем ближе $\varphi\left(\vec{\mathbf{X}}\right)$ к линейной функции и чем меньше максимальная из дисперсий σ_j^2.

Пусть между случайными векторами $\vec{\mathbf{X}}$ и $\vec{\mathbf{Y}}$ существует линейная зависимость вида

$$\vec{\mathbf{Y}} = \hat{\mathrm{A}}\,\vec{\mathbf{X}} + \vec{\mathbf{C}}\,, \tag{1.29}$$

где $\hat{\mathrm{A}}$ — матрица преобразования полного ранга (ранг квадратной матрицы $\hat{\mathrm{A}}$ размера $(m \times m)$ равен m); $\vec{\mathbf{C}}$ — вектор-столбец свободных членов. В этом случае формулы (1.28) являются точными, то есть

$$\vec{\boldsymbol{\eta}} = \hat{\mathrm{A}}\,\vec{\boldsymbol{\xi}} + \hat{\mathrm{C}}\,, \tag{1.30}$$

$$\hat{\Sigma}_{\vec{\mathbf{Y}}} = \hat{\mathrm{A}}\,\hat{\Sigma}_{\vec{\mathbf{X}}}\,\hat{\mathrm{A}}^T. \tag{1.31}$$

Для нахождения функции распределения случайного вектора $\vec{\mathbf{Y}}$ можно использовать ту же формулу, что и в одномерном случае:

$$F_{\vec{\mathbf{Y}}}\left(\vec{\mathbf{Y}}\right) = F_{\vec{\mathbf{X}}}\left(\varphi^{-1}\left(\vec{\mathbf{Y}}\right)\right), \tag{1.32}$$

где $\varphi^{-1}\left(\vec{\mathbf{Y}}\right)$ — функция, обратная к $\varphi\left(\vec{\mathbf{Y}}\right)$. Плотность распределения вектора $\vec{\mathbf{Y}} = \varphi\left(\vec{\mathbf{X}}\right)$ может быть найдена по формуле

$$f_{\vec{\mathbf{Y}}}\left(\vec{\mathbf{Y}}\right) = f_{\vec{\mathbf{X}}}\left(\varphi^{-1}\left(\vec{\mathbf{Y}}\right)\right)\left|\hat{\mathrm{J}}\left(\vec{\mathbf{Y}}\right)\right|, \tag{1.33}$$

где $\left|\hat{\mathrm{J}}\left(\vec{\mathbf{Y}}\right)\right|$ — якобиан преобразования, взятый со знаком плюс. Две последние формулы дают верный результат только при условии однозначности обратной функции $\varphi^{-1}\left(\vec{\mathbf{Y}}\right)$. В общем случае приходится применять эти формулы для отдельных областей однозначности функции φ и использовать теорему сложения вероятностей.

Для примера приведем выражение для плотности распределения линейной функции от случайного вектора $\vec{\mathbf{X}}$ (1.19):

$$f_{\vec{\mathbf{Y}}}\left(\vec{\mathbf{Y}}\right) = \frac{f_{\vec{\mathbf{X}}}\left(\hat{\mathrm{A}}^{-1}\left(\vec{\mathbf{Y}} - \vec{\mathbf{C}}\right)\right)}{\left\|\hat{\mathrm{A}}\right\|}\,. \tag{1.34}$$

Другим важным примером является линейное преобразование вида

$$Z = \sum_{j=1}^{m} x_j \tag{1.35}$$

в случае статистически независимых координат вектора $\vec{\mathbf{X}}$. Здесь уже $m \neq n$, так как Z — скаляр.

Для двумерного вектора $(x,\ y)$ плотность распределения случайной величины $Z = x + y$ выражается *сверткой маргинальных плотностей слагаемых* [8]:

$$\begin{aligned} f(Z) &= \int\limits_{-\infty}^{+\infty} f_1(x) f_1(Z - x)\, dx = \\ &= \int\limits_{-\infty}^{+\infty} f_1(Z - y) f_1(y)\, dy\,. \end{aligned} \tag{1.36}$$

При размерности вектора $\vec{\mathbf{X}}$ более двух, формулу (1.36) приходится применять многократно. Важность этого примера видна, например, из того факта, что обычно из наблюдений мы получаем не распределение исследуемого параметра $\vec{\mathbf{X}}$, которым могут быть, например, металличность $[\mathrm{Fe/H}]$, остаточная лучевая скорость в кинематических задачах и т. д., а распределение $\vec{\mathbf{X}} + \vec{\boldsymbol{\epsilon}}$, где $\vec{\boldsymbol{\epsilon}}$ — ошибка определения этого параметра.

Чтобы получить искомое распределение параметра, следует решить уравнение свертки, ибо наблюдаемое распределение есть, как сказано выше, свертка истинного распределения параметра и распределения его ошибок.

Еще одним линейным преобразованием случайного вектора является его *приведение к вектору с некоррелированными компонентами*. После такого преобразования ковариационная матрица нового вектора становится диагональной. Наиболее известным из подобного рода преобразований является преобразование вида

$$\vec{\mathbf{Z}} = \hat{\mathrm{V}}^T \left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}} \right), \tag{1.37}$$

где $\vec{\mathbf{Z}}$ — преобразованный случайный вектор; $\hat{\mathrm{V}}$ — $(m \times m)$-матрица преобразования, столбцы которой являются собственными векторами матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$. Элементы главной диагонали ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{Z}}}$ представляют собой собственные значения λ_j ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$, являющиеся дисперсиями компонент вектора $\vec{\mathbf{Z}}$. При этом $\hat{\mathrm{V}}^T\hat{\mathrm{V}} = \hat{\mathrm{I}}$.

Обратное преобразование

$$\left(\hat{\mathrm{V}}^T\right)^{-1} \vec{\mathbf{Z}} = \left(\hat{\mathrm{V}}^T\right)^{-1} \hat{\mathrm{V}}^T \left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}} \right), \tag{1.38}$$

$$\left(\hat{\mathrm{V}}^{-1}\right)^T \vec{\mathbf{Z}} = \vec{\mathbf{X}} - \vec{\boldsymbol{\xi}} \quad \left[\hat{\mathrm{V}}^T\hat{\mathrm{V}} = \hat{\mathrm{I}} \;\Rightarrow\; \hat{\mathrm{V}}^T = \hat{\mathrm{V}}^{-1} \right], \tag{1.39}$$

$$\left(\hat{\mathrm{V}}^T\right)^T \vec{\mathbf{Z}} = \vec{\mathbf{X}} - \vec{\boldsymbol{\xi}}, \tag{1.40}$$

$$\vec{\mathbf{X}} = \vec{\boldsymbol{\xi}} + \hat{\mathrm{V}}\vec{\mathbf{Z}}, \tag{1.41}$$

$$\vec{\mathbf{X}}^T = \vec{\boldsymbol{\xi}}^T + \vec{\mathbf{Z}}^T\hat{\mathrm{V}}^T = \vec{\boldsymbol{\xi}}^T + \sum_{k=1}^{m} z_k \vec{\mathbf{V}}_k, \tag{1.42}$$

где $\vec{\mathbf{V}}_k$ — k-й собственный вектор матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$ — является раз-

ложением вектора $\vec{\mathbf{X}}$ по собственным векторам его ковариационной матрицы (1.42).

Случайные величины z_k, $\vec{\mathbf{V}}_k$ называют **главными компонентами** случайного вектора $\vec{\mathbf{X}}$, так как среди проекций вектора $\vec{\mathbf{X}}$ на все возможные направления максимальную дисперсию, равную максимальному собственному значению (пусть это будет λ_i), имеет проекция на собственный вектор, соответствующий этому собственному значению $\vec{\mathbf{V}}_i$, и т. д.

Геометрически преобразование (1.37) сводится к переносу начала координат в точку $\vec{\boldsymbol{\xi}}$ и повороту координатных осей таким образом, чтобы оси новой системы координат совпали с главными осями эллипсоида рассеяния (1.17).

В заключение отметим, что все линейные преобразования не изменяют закон распределения случайного вектора, изменяются лишь числовые параметры закона.

В прил. 1 приведены численные примеры получения ковариационных и корреляционных матриц и коэффициентов корреляции.

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ МНОГОМЕРНЫХ РАСПРЕДЕЛЕНИЙ

2.1. Параметры распределений

Рассматриваемые в теории вероятностей законы распределения случайных величин являются лишь математическими моделями, более или менее адекватно аппроксимирующими реальные случайные явления. Однако, если говорить не о законах, а только о параметрах распределений, таких как $\vec{\boldsymbol{\xi}}$, $\hat{\Sigma}$, $\hat{\mathrm{P}}$ — вектор математических ожиданий, ковариационная или корреляционная матрицы, то эти понятия не являются модельными, а присущи, независимо от закона распределения, всем случайным величинам, для которых эти параметры существуют. Поэтому задача оценивания неизвестных характеристик реальных случайных величин является основной и важнейшей в математической статистике.

Очевидно, что *точные значения параметров распределений найти невозможно*, поскольку невозможно провести бесконечно большое число наблюдений над поведением случайной величины в идентичных условиях.

Определение 2.1. *Бесконечное число результатов измерений случайной величины, сделанных в идентичных условиях, в математической статистике называют* ***генеральной совокупностью****, а конечный ряд полученных в эксперименте значений реализаций случайной величины* $\vec{\mathbf{X}}$*, содержащий* n *результатов, —* ***выборкой объема*** n *из генеральной совокупности.*

Следует сделать оговорку, что иногда под генеральной совокупностью понимают реальную конечную совокупность однородных в каком-то смысле объектов (звезд в Галактике, людей на Земле), однако это не меняет сути дела, которая заключается в том, чтобы по относительно небольшой выборке сделать наиболее достоверные выводы о свойствах и характеристиках всей совокупности.

Отметим, во избежание путаницы, что под параметрами распределения иногда понимают не математическое ожидание, ковариационную матрицу и плотность вероятности, но параметры, входящие в формулу плотности вероятности. Если в случае нормального распределения среднее (в многомерном случае — вектор математических ожиданий) и дисперсия (ковариационная матрица) прямо входят в выражение для плотности распределения, то для других распределений это не так. Например, распределение Коши (для простоты мы берем одномерный случай) с плотностью вероятности

$$f(x,\ \lambda,\ \mu) = \frac{\lambda}{\pi\left(\lambda^2 + (x-\mu)^2\right)} \tag{2.1}$$

вообще не имеет моментов четных порядков (соответствующие интегралы расходятся), и, в частности, параметр μ является

модой и медианой распределения, а λ не связана напрямую с дисперсией. Поэтому то, что понимается под определяемыми параметрами, будет конкретизироваться в каждом отдельном случае. (Следует отметить, что для симметричных выборочных распределений распределение Коши можно удобно использовать при оценивании моды выборочного распределения.)

2.2. Выборка и оценка

Выборка из некоторой генеральной совокупности представляет собой набор числовых значений — реализаций изучаемой случайной величины $\vec{\mathbf{X}}$, полученных по n наблюдениям. *Если случайная величина одномерна*, то элементом выборки является число $x_k,\ k = 1,\ 2,\ \ldots,\ n$, а саму выборку можно рассматривать как случайный вектор $\vec{\mathbf{X}}^T = (x_1,\ \ldots,\ x_n)$, компоненты которого меняются от выборки к выборке случайным образом.

Определение 2.2. *Для многомерной случайной величины элементом выборки будет выборочный вектор*

$$\vec{\mathbf{X}}_i^T = (x_{i1},\ \ldots,\ x_{in})\,, \tag{2.2}$$

где первый индекс i — номер компоненты случайного вектора $\vec{\mathbf{X}}$ и $i = 1,\ 2,\ \ldots,\ m$; второй индекс — номер наблюдения в выборке размером n. Саму выборку можно представить в виде $(n \times m)$-матрицы, строки которой являются элементами выборки. Эту матрицу часто называют ***матрицей данных*** $\hat{\mathrm{M}}$.

Важнейшим требованием к выборке является то, что она *должна быть извлечена из одной неизменной генеральной совокупности*, что на практике выполнить достаточно трудно, так как любое случайное явление обычно подвержено влиянию неслучайных факторов. Для исключения этого влияния или для сведения его к случайному прибегают к **рандомизации**, то есть проводят эксперименты и наблюдения в разное время суток, разные сезоны, на разных приборах и обсерваториях. Желательно также, чтобы выборка состояла из статистически независимых элементов, что тоже достигается рандомизацией. В дальнейшем речь пойдет только о независимых выборках из одной генеральной совокупности.

Определение 2.3. *Будем называть* ***оцениванием*** *процедуру получения численного значения некоторого параметра* Θ*;* ***оценкой*** $\breve{\Theta}$ *— само это численное значение;* ***статистикой*** *— формулу для получения оценки и вообще любую формулу, выражающую функцию от случайных величин. Если рассматривать элементы выборки как независимые одинаково распределенные случайные величины, то оценка будет случайной величиной, имеющей свой закон распределения.*

Как функция выборки оценка $\breve{\Theta}$ неизвестного параметра Θ является случайной величиной, распределение которой полностью определяется распределением генеральной совокупности. Оценка $\breve{\Theta}$ должна обладать определенными свойствами. Она должна быть

— **состоятельной**, то есть должна сходиться по вероятности к величине оцениваемого параметра Θ при $n \to \infty$;

— **несмещенной**, то есть ее математическое ожидание $\hat{\mathrm{E}}\left(\breve{\Theta}\right)$

должно быть равно Θ;

— **эффективной**, то есть обладать наименьшей дисперсией среди всех возможных оценок данного параметра (условие желательное, но не обязательное).

2.3. Точечные и интервальные оценки для математического ожидания и ковариационной матрицы

Определение 2.4. *Для одномерных случайных величин наилучшими с точки зрения перечисленных выше требований оценками математического ожидания и дисперсии являются* ***среднее арифметическое из элементов выборки*** *и* ***выборочная несмещенная дисперсия****. Многомерными аналогами этих выражений являются* ***вектор средних значений*** *и* ***выборочная ковариационная матрица****:*

$$\langle \vec{\mathbf{X}} \rangle = \frac{1}{n} \hat{\mathrm{M}}^T \vec{\mathbf{I}}_n \,, \tag{2.3}$$

$$\hat{\mathrm{S}}_{\vec{\mathbf{X}}} = \frac{1}{n-1} \left(\hat{\mathrm{M}}^T \hat{\mathrm{M}} - n \langle \vec{\mathbf{X}} \rangle \langle \vec{\mathbf{X}} \rangle^T \right) , \tag{2.4}$$

где $\hat{\mathrm{M}}$ *— матрица данных;* $\vec{\mathbf{I}}_n$ *— вектор с* n *компонентами, равными единице. Угловыми скобками обозначены соответствующие средние значения, так что элементами вектора* $\langle \vec{\mathbf{X}} \rangle$ *и выборочной ковариационной матрицы* $\hat{\mathrm{S}}_{\vec{\mathbf{X}}}$ *являются со-*

ответственно

$$\langle x_k \rangle = \frac{1}{n} \sum_{i=1}^{n} x_{ik}, \tag{2.5}$$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \langle x_j \rangle \right) \left(x_{ik} - \langle x_k \rangle \right) =$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_{ij}\, x_{ik} - \frac{1}{n} \sum_{i=1}^{n} x_{ij} \sum_{i=1}^{n} x_{ik} \right), \tag{2.6}$$

где суммирование ведется по элементам выборки.

Последний вариант вычисления элементов выборочной ковариационной матрицы удобен для вычисления на компьютерах тем, что вычисления средних и ковариаций ведутся в одном цикле. Попутно отметим, что элементами **выборочной корреляционной матрицы** являются величины

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}\, s_{kk}}}\,. \tag{2.7}$$

Определение 2.5. *Оценки вида* (2.3), (2.4) *называются* ***точечными****, так как дают одно числовое значение оценки (точку в пространстве соответствующей размерности), не давая никакой информации о погрешности оценки.*

Определение 2.6. ***Интервальные оценки*** *задают область со случайными границами, внутри которой с заданной вероятностью γ может находиться истинное значение оцениваемого параметра. Такая область называется* ***доверительной областью****, а соответствующая вероятность —* ***доверительной вероятностью****.*

Построение интервальной оценки сводится, таким образом, к отысканию доверительной области *по заданной доверительной вероятности*, обычно близкой к единице. Часто ее берут равной 95 %, иногда 90 и 99 %. Естественно, для решения этой задачи необходимо знать распределение оценки, которое, как отмечалось выше, полностью определяется распределением генеральной совокупности.

Определение 2.7. *Доверительная область для оценки математического ожидания $\vec{\boldsymbol{\xi}}$ нормально распределенного случайного вектора $\vec{\mathbf{X}}$ для доверительной вероятности γ задается формулой [9]*

$$\left(\langle\vec{\mathbf{X}}\rangle - \vec{\boldsymbol{\xi}}\right)^T \hat{\mathrm{S}}^{-1}_{\vec{\mathbf{X}}} \left(\langle\vec{\mathbf{X}}\rangle - \vec{\boldsymbol{\xi}}\right) \leqslant \frac{m\,(n-1)}{n\,(n-m)}\, F_\alpha(m,\, n-m)\,, \quad (2.8)$$

где F_α — α-процентная точка распределения Фишера*, определяемая из таблиц по аргументам $\alpha = 1-\gamma$, m и $n-m$, где m — размерность случайного вектора $\vec{\mathbf{X}}$; n — число элементов выборки. Геометрически область, задаваемая выражением* (2.8)*, представляет собой m-мерный гиперэллипсоид с центром в точке, задаваемой вектором $\langle\vec{\mathbf{X}}\rangle$.*

Доверительная область для ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$ нормально распределенного случайного вектора $\vec{\mathbf{X}}$ строится в $m\,(m+1)/2$-мерном пространстве с помощью *распределения Уишарта (J. Wishart)*, которое в определенном смысле является многомерным обобщением χ^2 распределения. Справедлива следующая теорема [10, 11].

Теорема 2.1. *Пусть* $\vec{\mathbf{X}}_1, \ldots, \vec{\mathbf{X}}_n$ *— независимые случайные векторы, распределенные по нормальному закону с математическим ожиданием* $\vec{\boldsymbol{\xi}}$ *и ковариационной матрицей* $\hat{\Sigma}_{\vec{\mathbf{X}}}$*, причем* $n \geqslant m+1$*, где* m *— размерность векторов* $\vec{\mathbf{X}}_i$*. Тогда матрица вида (выборочная ковариационная матрица)*

$$\hat{\mathrm{S}}_{\vec{\mathbf{X}}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\vec{\mathbf{X}}_i - \langle \vec{\mathbf{X}} \rangle \right) \left(\vec{\mathbf{X}}_i - \langle \vec{\mathbf{X}} \rangle \right)^T \tag{2.9}$$

имеет распределение $W\left((n-1)^{-1} \hat{\Sigma}_{\vec{\mathbf{X}}},\ n-1 \right)$ *— распределение Уишарта с ковариационной матрицей* $(n-1)^{-1} \hat{\Sigma}_{\vec{\mathbf{X}}}$ *и* $(n-1)$ *степенями свободы.*

2.4. Проверка гипотез о параметрах распределения

Как уже отмечалось, *по выборке никаких категорических утверждений о параметрах распределения сделать невозможно*, поэтому приходится ограничиваться более или менее правдоподобными предложениями — гипотезами — о значениях этих параметров.

Цель статистической проверки гипотез — установить, противоречит ли выдвинутое предположение экспериментальным данным и с какой степенью достоверности.

Определение 2.8. ***Статистической гипотезой*** *принято считать любое предположение о законе распределения случайной величины или о значениях параметров закона распределения.*

Задача проверки гипотез решается на основе принципа **практической невозможности**, согласно которому событие с малой вероятностью (например, $< 5\ \%$) в конкретных условиях эксперимента может считаться практически невозможным. Вероятность события вычисляется на основе выдвинутой гипотезы о законе распределения и его параметрах, поэтому если при испытании практически невозможное событие все-таки произойдет, то наши предпосылки вряд ли верны и гипотезу о законе распределения и его параметрах следует отвергнуть. Вероятность практически невозможного события называют **уровнем значимости** α и выбирают ее исходя из конкретных условий эксперимента. Чаще всего используют значения 0.01 и 0.05.

Процедура проверки гипотез о законе и параметрах распределения сводится к следующим шагам:

1. Выбор уровня значимости.
2. Вычисление статистики критерия, являющейся функцией выборки — случайной величиной.
3. Построение критической области, вероятность попадания значения статистики критерия в которую равна α.

Для этого, как и при построении доверительной области, необходимо знать распределение статистики критерия, но в данном случае *это распределение задано гипотезой.* Если значение критерия попадает в критическую область, то гипотеза отвергается, при этом вероятность отвергнуть правильную гипотезу не превосходит уровня значимости α. В противном случае говорят, что нет оснований отвергнуть гипотезу.

Высказанное предположение, которое подлежит проверке, обозначается H_0 и называется **нулевой гипотезой**. Наряду с нулевой в рассмотрение вводится и противоречащая альтернативная гипотеза H_1. Гипотезы о значении параметра распределения записывают символически в виде H_0: $\breve{\Theta} = \breve{\Theta}_0$, где $\breve{\Theta}_0$ — предполагаемое значение параметра $\breve{\Theta}$. Например, гипотеза о равенстве нулю средней пространственной скорости группы звезд может быть записана как H_0: $\hat{\mathrm{E}}(V) = 0$.

Построение критической области математически почти идентично построению доверительной области. Например, для гипотезы H_0: $\vec{\boldsymbol{\xi}} = \vec{\boldsymbol{\xi}}_0$ в случае нормально распределенной генеральной совокупности уравнением критической области будет соотношение (2.8), но с обратным знаком неравенства, а статистикой критерия — левая часть (2.8), где вместо $\vec{\boldsymbol{\xi}}$ следует поставить $\vec{\boldsymbol{\xi}}_0$.

Проверка гипотезы H_0: $\vec{\boldsymbol{\xi}} = \vec{\boldsymbol{\eta}}$ о равенстве математических ожиданий двух m-мерных нормально распределенных случайных векторов $\vec{\mathbf{X}}$ и $\vec{\mathbf{Y}}$, например о равенстве средних пространственных скоростей двух центроидов звезд, проводится с помощью статистики $\left(\langle\vec{\mathbf{X}}\rangle - \langle\vec{\mathbf{Y}}\rangle\right)^T \hat{\mathrm{H}}^{-1} \left(\langle\vec{\mathbf{X}}\rangle - \langle\vec{\mathbf{Y}}\rangle\right)$, а критическая область определяется неравенством [11]

$$\begin{aligned}\left(\langle\vec{\mathbf{X}}\rangle - \langle\vec{\mathbf{Y}}\rangle\right)^T \hat{\mathrm{H}}^{-1} \left(\langle\vec{\mathbf{X}}\rangle - \langle\vec{\mathbf{Y}}\rangle\right) > \\ > \frac{m\,\left(n_{\vec{\mathbf{X}}} + n_{\vec{\mathbf{Y}}} - 2\right)}{n_{\vec{\mathbf{X}}}\, n_{\vec{\mathbf{Y}}}\left(n_{\vec{\mathbf{X}}} + n_{\vec{\mathbf{Y}}} - m - 1\right)} \times \\ \times\, F_\alpha(m,\; n_{\vec{\mathbf{X}}} + n_{\vec{\mathbf{Y}}} - m - 1)\,,\end{aligned} \tag{2.10}$$

где $\hat{\mathrm{H}} = \left(n_{\vec{\mathbf{X}}} - 1\right) \hat{\mathrm{S}}_{\vec{\mathbf{X}}} + \left(n_{\vec{\mathbf{Y}}} - 1\right) \hat{\mathrm{S}}_{\vec{\mathbf{Y}}}$; $\hat{\mathrm{S}}_{\vec{\mathbf{X}}}$ и $\hat{\mathrm{S}}_{\vec{\mathbf{Y}}}$ — выборочные ковариационные матрицы; $n_{\vec{\mathbf{X}}}$ и $n_{\vec{\mathbf{Y}}}$ — объемы выборок.

Вариантом этой гипотезы является выявление аномальных элементов выборки, например, грубых промахов в результатах измерений или при подготовке данных к дальнейшей обработке. При этом промахи не обязательно вызываются сбоем приборов или грубыми ошибками наблюдателей. Часто промахи появляются при включении в выборку элементов из другой генеральной совокупности. Как примеры таких ситуаций можно рассматривать включение звезд фона в число членов изучаемого звездного скопления.

Проверка на «вылет». Допустим, что в выборке объема $n+1$ из генеральной совокупности $\vec{\mathbf{X}}$ заметно отклонившимся от среднего значения оказался $(n+1)$-й элемент. Возникает предположение, что этот элемент является аномальным для совокупности $\vec{\mathbf{X}}$, то есть что он извлечен из другой генеральной совокупности $\vec{\mathbf{Y}}$. Проверить это предположение можно при помощи выражения (2.10), положив в нем $n_{\vec{\mathbf{X}}} = n$ и $n_{\vec{\mathbf{Y}}} = 1$. Критическая область при этом определится неравенством

$$\left(\langle\vec{\mathbf{X}}\rangle - \vec{\mathbf{Y}}\right)^T \hat{\mathrm{H}}^{-1} \left(\langle\vec{\mathbf{X}}\rangle - \vec{\mathbf{Y}}\right) > \\ > \frac{m\,(n+1)}{n\,(n-m)}\, F_\alpha(m,\ n-m)\,, \tag{2.11}$$

где $\langle\vec{\mathbf{X}}\rangle$ — вектор среднего по выборке $n_{\vec{\mathbf{X}}}$, то есть без $(n+1)$-го элемента; $\vec{\mathbf{Y}} = \vec{\mathbf{X}}_{n+1}$; $\hat{\mathrm{H}} = \left(n_{\vec{\mathbf{X}}} - 1\right) \hat{\mathrm{S}}_{\vec{\mathbf{X}}}$. Аналогично проводится выявление нескольких подозрительных на аномальность элементов выборки.

Особо следует подчеркнуть, что формулы (2.5)—(2.11) справедливы только для случайных векторов с нормальным распределением.

2.5. Оценивание параметров угловых случайных величин

В астрономии и геодезии как ни в каких других науках широко используются угловые величины. Это связано, в частности, с понятием небесной сферы, на которой применимы *только угловые величины*. Случайными угловые величины оказываются либо в результате влияния погрешностей измерений, либо они являются координатами некоторого случайно выбранного объекта изучаемой совокупности. К угловым величинам можно относить также временные величины в периодических процессах. Во всех указанных случаях размерность углового случайного вектора не превосходит двойки, а сами значения углов обычно заключены в пределах либо $[\,0,\ 2\pi)$, либо $[\,-\pi/2,\ +\pi/2\,]$. Для удобства изложения будем рассматривать только первый вариант.

Использование методов оценивания параметров распределения, принятых для «обычных» случайных величин, *не представляется возможным*, так как это может приводить к грубым ошибкам и артефактам. Это является следствием того, что угловые величины определяются на окружности, а не на разомкнутой прямой (линии) и *терпят разрыв* в окрестности начального направления. Например, использование обычного

метода оценки математического ожидания и дисперсии для угловой величины, которая реализуется вокруг 0°, приведет к совершенно неправдоподобным результатам.

Любой угол на плоскости можно представить точкой на окружности единичного радиуса с центром в вершине угла, совпадающей с началом системы координат (рис. 2.1).

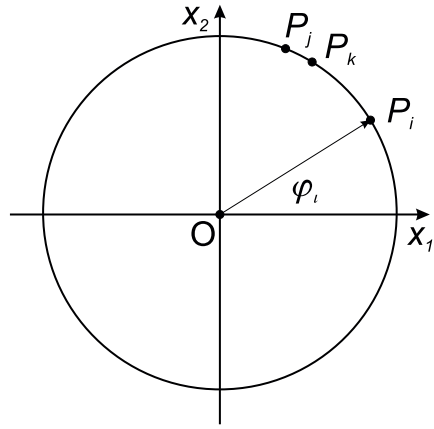

Рис. 2.1. Случайная угловая величина в векторном представлении

Пусть P_i — точка на окружности единичного радиуса, задаваемая углом φ_i, где $i = 1, \ldots, n$; n — количество элементов выборки. Тогда оценка математического ожидания для угловой переменной может быть определена следующим образом [12].

Определение 2.9. ***Выборочным круговым средним направлением*** *для набора углов φ_i является направление векторной суммы векторов $\overrightarrow{OP_i}$ (рис. 2.1).*

В декартовой системе координат точки P_i имеют координаты $(\cos\varphi_i,\ \sin\varphi_i)$. Если все точки выборки равноценны, то координаты $(C_1,\ C_2)$ центра масс системы точек $\{P_i\}$, направление на который из начала системы координат совпадает с направлением суммы векторов $\overrightarrow{OP}_i$, вычисляются как

$$C_1 = \frac{1}{n}\sum_{i=1}^{n}\cos\varphi_i\,, \tag{2.12}$$

$$C_2 = \frac{1}{n}\sum_{i=1}^{n}\sin\varphi_i\,, \tag{2.13}$$

причем

$$r_C = \sqrt{C_1^2 + C_2^2} \tag{2.14}$$

есть модуль (длина) вектора центра масс. Длина суммы векторов $\sum_{i=1}^{n}\overrightarrow{OP}_i$ вычисляется как $R = n\,r_C$.

С другой стороны, координаты вектора центра масс системы точек $\{P_i\}$ могут быть представлены следующим образом:

$$C_1 = r_C\,\cos\varphi_m\,, \tag{2.15}$$

$$C_2 = r_C\,\sin\varphi_m\,, \tag{2.16}$$

где φ_m — направление суммы векторов $\sum_{i=1}^{n}\overrightarrow{OP}_i$, которое является искомым **выборочным круговым средним направлением** φ_m. В случае когда $r_C = 0$, выборочное круговое среднее направление *не определяется*.

Поскольку все векторы P_i имеют *одинаковую единичную длину*, то значение r_C характеризует взаимное расположение (разброс) векторов P_i на окружности единичного радиуса. Например, если точки P_i равномерно распределены по окружно-

сти, то значение r_C стремится к нулю. Если же все P_i совпадают друг с другом (нулевой разброс), то значение r_C равно единице.

Для оценки разброса случайной угловой величины относительно выборочного кругового среднего направления необходимо ввести понятие **отклонения угловой величины от заданного направления**.

Определение 2.10. *Отклонение направления $\overrightarrow{OP_i}$ от направления, задаваемого углом Φ_0, определим в общем случае неограниченной области определения угловых величин как*

$$\begin{aligned} \Delta\varphi_i &= \min\left\{(\varphi_i - \Phi_0)\text{`}\,,\ 2\pi - (\varphi_i - \Phi_0)\text{`}\right\} = \\ &= \pi - \left|\pi - (\varphi_i - \Phi_0)\text{`}\right|\,, \\ \Delta\varphi_i &\in [\,0,\ \pi\,]\,, \end{aligned} \tag{2.17}$$

где $(\varphi_i - \Phi_0)$` — дробная часть разности $(\varphi_i - \Phi_0)$ по модулю 2π:

$$(\varphi_i - \Phi_0)\text{`} \equiv (\varphi_i - \Phi_0) - 2\pi \left[\frac{\varphi_i - \Phi_0}{2\pi}\right]\,, \tag{2.18}$$

где $[\cdots]$ — оператор, вычисляющий наибольшее целое число, не превышающее выражение внутри квадратных скобок.

Функция вида $1 - \cos\Delta\varphi_i = 2\sin^2(\Delta\varphi_i/2)$ является монотонно возрастающей на $[\,0,\ \pi\,]$.

Определение 2.11. ***Выборочной характеристикой рассеяния*** *направлений φ_i относительно направления Φ_0 называется функция $V(\Phi_0)$:*

$$V(\Phi_0) = \frac{1}{n}\sum_{i=1}^{n}(1-\cos\Delta\varphi_i) = \\ = \frac{2}{n}\sum_{i=1}^{n}\sin^2\frac{\Delta\varphi_i}{2}\,. \tag{2.19}$$

Для выборочной характеристики рассеяния (2.19) и выборочного кругового среднего направления φ_m выполняется следующее тождество:

$$V(\Phi_0) \equiv V(\varphi_m) + 2\,r_C\sin^2\left(\frac{\varphi_m-\Phi_0}{2}\right), \quad r>0\,, \tag{2.20}$$

из чего следует, что выборочная характеристика рассеяния, полученная относительно выборочного кругового среднего направления φ_m, является *минимальной выборочной характеристикой рассеяния из всех возможных*. Аналогичным свойством обладает и дисперсия как оценка разброса случайной величины относительно среднего арифметического.

Определение 2.12. ***Выборочной круговой дисперсией*** *направлений φ_i будем называть величину*

$$
\begin{aligned}
V(\varphi_m) &= \frac{1}{n}\sum_{i=1}^{n}(1-\cos\Delta\varphi_i) = \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1-\cos\left(\varphi_i-\varphi_m\right)'\right) = \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\cos(\varphi_i-\varphi_m)) = \\
&= 1-\frac{1}{n}\sum_{i=1}^{n}\cos(\varphi_i-\varphi_m) = 1-r_C\,, \\
V(\varphi_m) &\in [\,0,\ 1\,]\,.
\end{aligned}
\tag{2.21}
$$

Выборочной результирующей длиной r_C *называется величина*

$$
r_C = \frac{R}{n} = \frac{\sum\limits_{i=1}^{n}\overrightarrow{OP}_i}{n} = \sqrt{C_1^2+C_2^2}\,. \tag{2.22}
$$

Выборочная круговая дисперсия *не зависит от выбора начала отсчета углов.*

Для оценки рассеяния угла относительно выборочного среднего направления также может быть введена величина, оценивающая разброс угла непосредственно в радианах или градусах, в отличие от выборочной круговой дисперсии направления, которая имеет размерность линейной величины.

Определение 2.13. ***Выборочным круговым стандартным отклонением*** *[13] называется величина*

$$
\begin{aligned}
& s_\varphi = \sqrt{-\,2\ln\left(1 - V(\varphi_m)\right)} = \sqrt{-2\ln r_C}\,, \\
& s_\varphi \in [\,0,\ +\infty\,]\ .
\end{aligned}
\tag{2.23}
$$

Для малых значений $V(\varphi_m)$ выполняется соотношение, аналогичное таковому для обычным образом определенных дисперсии и стандартного отклонения: $s_\varphi \approx \sqrt{2\,V(\varphi_m)}$.

Все сказанное об оценках параметров случайных углов может быть обобщено на двумерный случайный угловой вектор, что необходимо, так как положение точки на небесной сфере задается двумя углами, например, галактической долготой l и широтой b. Вектор положения $\vec{\mathbf{X}}$ на небесной сфере (обычный пространственный вектор, а не угловой) при этом будет трехмерным: $\vec{\mathbf{X}} = (\cos l \cos b,\, \sin l \cos b,\, \sin b)$. При необходимости анализировать положение случайной точки на небесной сфере и оценивать моменты такой случайной величины следует вычислять величины $\Delta\Phi_i$ (углы между векторами $\vec{\mathbf{X}}_i$ и $\langle\vec{\mathbf{X}}\rangle$) по формулам сферической тригонометрии как центральные углы в соответствующих плоскостях больших кругов. Угловое среднее положение на сфере может быть определено исходя из того, что сферические системы координат являются ортогональными, и процедура определения выборочного кругового среднего значения применима к каждой из сферических координат вектора на небесной сфере. Дополнительная информация о статистической обработке положений объектов в астрономии есть в [14, 15].

3. ОЦЕНИВАНИЕ ПЛОТНОСТИ И ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

3.1. Распределения Пирсона

Изложенные в предыдущем разделе методы оценивания параметров распределения в принципе дают возможность оценивать и сами распределения, поскольку, как утверждается в теории вероятностей, *полный набор моментов случайной величины однозначно определяет ее распределение*, а моменты можно оценить по выборке. Однако путь этот в общем случае бесперспективен, так как восстановить аналитический вид функции распределения по набору моментов невозможно. Кроме того, моменты высоких порядков оцениваются по выборкам настолько неточно, что говорить об их использовании для оценивания функции распределения не приходится. Тем не менее оценки моментов распределения оказываются полезными в тех случаях, когда закон распределения известен с точностью до значений параметров. Например, если нам известно, что некоторая случайная величина распределена нормально, то оценивание всего двух параметров — математического ожидания и ковариационной матрицы — позволит оценить конкретный вид распределения [16].

Под **оцениванием распределений**, в сущности, понимаются две задачи. Первая — это выбор аналитического выражения для плотности или функции распределения на основе имеющейся в выборке информации, вторая — оценивание параметров выбранного или заранее известного распределения.

При выборе аналитического представления плотности распределения (для простоты изложения ограничимся пока плотностью распределения) на основе информации, содержащейся в выборке (а также и имеющейся априорной информации об изучаемом явлении), мы должны выбрать функцию из какого-либо семейства. В частности, таким семейством являются **распределения Пирсона**, для которого разработан метод выделения функции из семейства с использованием выборочных моментов.

Определение 3.1. ***Распределениями Пирсона*** *называются функции* $p = p(x)$*, удовлетворяющие дифференциальному уравнению вида*

$$\frac{dp(x)}{dx} = \frac{x - a}{b_0 + b_1 x + b_2 x^2}\, p(x)\,, \tag{3.1}$$

где параметры $a,\ b_0,\ b_1,\ b_2$ *— действительные числа. Графики зависимости* $p(x)$ *от* x *называются* ***кривыми Пирсона****.*

Распределения, являющиеся решениями уравнения (3.1), совпадают с предельными формами так называемого **гипергеометрического распределения**. Семейство кривых Пирсона составляют двенадцать типов и нормальное распределение. Выделяют следующие типы.

Тип I

$$p(x) = k\left(1 + \frac{x}{a_1}\right)^{m_1}\left(1 - \frac{x}{a_2}\right)^{m_2}, \qquad -a_1 \leqslant x \leqslant a_2, \quad m_1, m_2 > -1. \tag{3.2}$$

Частным случаем этого типа является *бета-распределение первого рода*.

Кривые первого типа разделяют на три группы в соответствии с поведением функции на интервале $[\,-a_1,\ a_2\,]$.

1. Кривая типа I. Реализуется при $m_1 > 0,\ m_2 > 0$. В этом случае значения функции $p(x)$ ограничены на интервале $[\,-a_1,\ a_2\,]$.
2. Кривая типа I(J). Реализуется в случае, если m_1 и m_2 имеют разные знаки. Значение функции $p(x)$ на одном из концов интервала $[\,-a_1,\ a_2\,]$ стремится к бесконечности.
3. Кривая типа I(U). Реализуется при $m_1 < 0,\ m_2 < 0$. Функция $p(x)$ достигает минимума на $(-a_1,\ a_2)$. На обоих краях интервала $[\,-a_1,\ a_2\,]$ функция $p(x)$ стремится к бесконечности.

Тип II

$$p(x) = k\left(1 + \frac{x}{a^2}\right)^{m}, \quad -a \leqslant x \leqslant a, \quad m \geqslant -1. \tag{3.3}$$

Частным случаем этого типа (при $m = 0$) является *равномерное распределение*.

Тип III

$$p(x) = k\left(1+\frac{x}{a}\right)^{-\mu x}, \qquad -a \leqslant x \leqslant \infty, \quad \mu > 0, \quad a > 0. \tag{3.4}$$

Частные случаи — *гамма-распределение*, *«хи-квадрат»-распределение*.

Тип IV

$$p(x) = k\left(1+\frac{x^2}{a^2}\right)^{-m} \exp\left(\mu \operatorname{arctg}\left(\frac{x}{a}\right)\right), \qquad -\infty \leqslant x \leqslant \infty, \quad a > 0, \quad \mu > 0. \tag{3.5}$$

Тип V

$$p(x) = k\, x^{-1} \exp\left(-\frac{\alpha}{x}\right), \qquad 0 \leqslant x < \infty, \quad \alpha > 0, \quad q > 1. \tag{3.6}$$

Этот тип сводится преобразованием к типу III.

Тип VI

$$p(x) = k\, x^{-q_1}(x-a)^{q_2}, \quad a \leqslant x < \infty, \quad q_1 > q_2 - 1. \tag{3.7}$$

Частные случаи — *бета-распределение второго рода*, а также F-распределение Фишера.

Тип VII

$$p(x) = k\left(1+\frac{x^2}{a^2}\right)^{-m}, \quad -\infty < x < \infty, \quad m > 0.5. \tag{3.8}$$

Частный случай — *распределение Стьюдента*.

Тип VIII

$$p(x) = k\left(1 + \frac{x}{a}\right)^{-m}, \quad -a \leqslant x \leqslant 0, \quad 0 \leqslant m \leqslant 1. \qquad (3.9)$$

Тип IX

$$p(x) = k\left(1 + \frac{x}{a}\right)^{m}, \quad -a \leqslant x \leqslant 0, \quad m > -1. \qquad (3.10)$$

Тип X

$$p(x) = k\exp\left(-\frac{x-m}{\sigma}\right), \quad m \leqslant x < \infty, \quad \sigma > 0. \qquad (3.11)$$

Это распределение называется *показательным*.

Тип XI

$$p(x) = k\,x^{-m}, \quad b \leqslant x < \infty, \quad m > 0. \qquad (3.12)$$

Распределение типа XI носит название *распределения Парето*, оно получило широкое распространение в задачах экономической статистики.

Тип XII

$$p(x) = \left(\frac{1 + \dfrac{x}{a_1}}{1 - \dfrac{x}{a_2}}\right)^{m}, \quad -a_1 \leqslant x \leqslant a_2, \quad |m| > 1. \qquad (3.13)$$

Этот тип соответствует варианту типа I.

Для практического использования распределений Пирсона важно следующее их свойство.

Теорема 3.1. *Всякая кривая Пирсона однозначно определяется своими первыми четырьмя моментами, если они конечны.*

Как показала практика применения кривых Пирсона, наиболее употребительными оказываются первые семь типов кривых.

Рассмотрим *процедуру выбора типа кривой*, приближающей выборочную плотность распределения. Центральные моменты по элементам выборки объема n оцениваются по формулам

$$\breve{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \langle x\rangle\right)^k , \tag{3.14}$$

где $k = 2$ для оценки дисперсии, $k = 3$ и $k = 4$ для центральных моментов третьего (асимметрия) и четвертого (эксцесс) порядка. Эти оценки являются *состоятельными, но смещенными.* Несмещенные оценки моментов вычисляются по формулам

$$\begin{aligned} m_2 &= \frac{n}{n-1}\,\breve{\mu}_2 , \\ m_3 &= \frac{n^2}{(n-1)(n-2)}\,\breve{\mu}_3 , \\ m_4 &= \frac{n(n^2-2n+3)\,\breve{\mu}_4 - 3n(2n-3)\,\breve{\mu}_2^2}{(n-1)(n-2)(n-3)} . \end{aligned} \tag{3.15}$$

Отметим, что при малых n смещения могут быть существенными, уменьшаясь с ростом объема выборки.

Введем **выборочные показатели асимметрии** $\breve{\beta}_1$ и **эксцесса** $\breve{\beta}_2$ распределения:

$$\breve{\beta}_1 = \frac{m_3}{m_2^{3/2}} , \tag{3.16}$$

$$\breve{\beta}_2 = \frac{m_4}{m_2^2} . \tag{3.17}$$

Далее можно двигаться двумя путями.

Первый вариант. Вычислим параметр κ, связанный с корнями знаменателя уравнения Пирсона (3.1):

$$\kappa = \frac{m_3^2 \left(m_4 + 3\right)^2}{4\left(4m_4 - 3m_3^2\right)\left(2m_4 - 3m_3^2 - 6\right)}\,. \tag{3.18}$$

В этом случае по вычисленному значению κ тип распределения Пирсона определяется из табл. 3.1.

Таблица 3.1

Выбор типа кривой Пирсона по значению параметра κ

Значение κ	Тип распределения Пирсона
$-\infty < \kappa < 0$	Тип I
$\kappa = 0, \quad m_3 = 0, \quad m_4 < 3$	Тип II
$\kappa = \pm\infty$	Тип III
$0 < \kappa < 1$	Тип IV
$\kappa = 1$	Тип V
$1 < \kappa < \infty$	Тип VI
$\kappa = 0, \quad m_3 = 0, \quad m_4 > 3$	Тип VII

Второй вариант определения типа кривой Пирсона — графический. Его применение, основанное на вычисляемых по выборке показателей асимметрии и эксцесса $\breve{\beta}_1$ и $\breve{\beta}_2$, ясно из рис. 3.1, с помощью которого и можно выбрать подходящую модель распределения. Здесь по оси абсцисс отложен показатель асимметрии, вычисляемый по формуле (3.16), а по оси ординат — эксцесс, вычисляемый по формуле (3.17).

Следует отметить, что изложенный здесь метод выбора модели распределения среди кривых Пирсона, а также и другие способы, такие как выбор *распределения Джонсона* или использование *ряда Эджворта*, основываются на выборочных

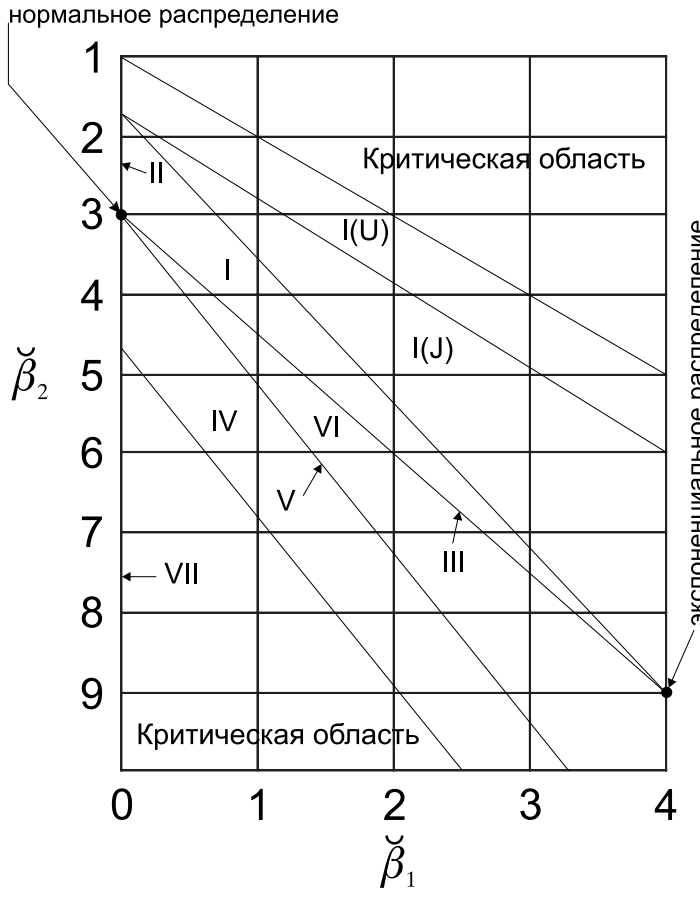

Рис. 3.1. Графический способ выбора типа кривой Пирсона. Кривой типа II соответствуют значения $\breve{\beta}_1 = 0$ и $\breve{\beta}_2 < 3$. Кривой типа VII соответствуют значения $\breve{\beta}_1 = 0$ и $\breve{\beta}_2 > 3$

оценках центральных моментов третьего и четвертого порядка, *очень чувствительных к промахам и вообще к структуре распределения точек на краях выборочного распределения.*

Иногда этот факт называют **неустойчивостью оценок**. Следует уяснить, что *по небольшим выборкам, содержащим порядка сотни значений, практически невозможно надежно оценить моменты третьего и четвертого порядка.*

Здесь уместно отметить, что часто удобно оценивать моменты распределения не по всей выборке, а *по гистограмме плотности распределения.* В этом случае заметно уменьшается объем вычислений. Однако при этом необходимо учитывать ошибку интервала, которая при большой ширине интервалов гистограммы может существенно сместить выборочные оценки моментов. Такие поправки, называемые **поправками Шеппарда**, для первых четырех моментов имеют следующий вид:

$$\begin{aligned} m'_1 &= m_1, \\ m'_2 &= m_2 - \frac{1}{12}\,h^2, \\ m'_3 &= m_3 - \frac{1}{4}\,m_1 h^2, \\ m'_4 &= m_4 - \frac{1}{2}\,m_2 h^2 + \frac{7}{240}\,h^4, \end{aligned} \tag{3.19}$$

где m'_i суть исправленные за ошибку интервала выборочные оценки моментов распределения; h — интервал (шаг) гистограммы. Как видно из выражений (3.19), *на оценку среднего значения ошибка интервала влияния не оказывает.*

3.2. Оценивание функции распределения

Определение 3.2. *Для получения оценки функции распределения $F(x)$ вероятность события $X < x$ заменяется на его* ***относительную частоту*** $n(x)/n$, *где $n(x)$ — число элементов выборки, меньших заданного значения аргумента x, а n — объем выборки. Таким образом,* ***выборочная функция распределения*** *определяется как*

$$\breve{F}(x) = \frac{n(x)}{n}\,. \tag{3.20}$$

График функции $\breve{F}(x)$ представляет собой неубывающую ступенчатую линию, имеющую n точек разрыва при значениях аргумента x_i.

Для построения доверительной области для $F(x)$ рассмотрим случайную величину

$$D = \max_x \left|\breve{F}(x) - F(x)\right|, \tag{3.21}$$

распределение которой при достаточно больших n не зависит от вида $F(x)$.

Доверительная область при заданной доверительной вероятности γ определяется неравенством

$$\left|\breve{F}(x) - F(x)\right| < \frac{d_\gamma}{\sqrt{n}}\,, \tag{3.22}$$

где d_γ — γ-процентная точка *распределения Колмогорова*. Распределение Колмогорова представляется в виде бесконечной суммы, поэтому его выражением на практике не пользуются,

хотя имеются ее таблицы. Полезно запомнить, что $d_{0.95} = 1.36$, а $d_{0.99} = 1.63$. Из (3.22) следует, что границы доверительной области в точности повторяют ход $\breve{F}(x)$, но смещены в направлении оси ординат вверх и вниз на величину $d_\gamma/\sqrt{n}$, однако, в силу определения $F(x)$, *не выходят за пределы интервала* $[\,0,\ 1\,]$.

Эмпирическая функция распределения в форме (3.20) неудобна при больших n, поэтому на практике используют ее упрощенный вид, получаемый при разбиении интервала, в котором лежат наши точки, на m подынтервалов:

$$\breve{F}(x_j) = \frac{n_j}{n}, \quad j = 1,\ 2,\ \ldots,\ m, \tag{3.23}$$

где n_j — число элементов выборки, меньших x_j. Величину m выбирают исходя из конкретной задачи исследования. Вопрос о выборе величины m не имеет строгого однозначного решения и во многом основывается на практическом опыте. Существует несколько широко применяемых способов выбора числа интервалов гистограммы. По формуле Брукса и Каррузера [7]

$$m = 5 \lg n\,. \tag{3.24}$$

Не менее широко распространено правило Стерджеса [17]

$$m = 1 + [\,\log_2 n\,]\,, \tag{3.25}$$

где $[\ldots]$ — операция взятия целой части числа. Интервалы между отдельными значениями x_j могут быть неравными друг другу, однако дальнейшая обработка существенно упростится, если они будут постоянными. Значения x_1 и x_m обычно выбирают равными наименьшему и наибольшему элемен-

там выборки соответственно. При построении эмпирической функции распределения удобно распределить элементы выборки по возрастанию, то есть представить в виде *вариационного ряда*. Для примера на рис. 3.2 показана функция распределения металличностей [Fe/H] шаровых скоплений нашей Галактики. Всего для построения функции использовано 135 оценок [Fe/H], построение проведено для десяти интервалов [Fe/H], штриховыми линиями показан 99 %-й доверительный интервал.

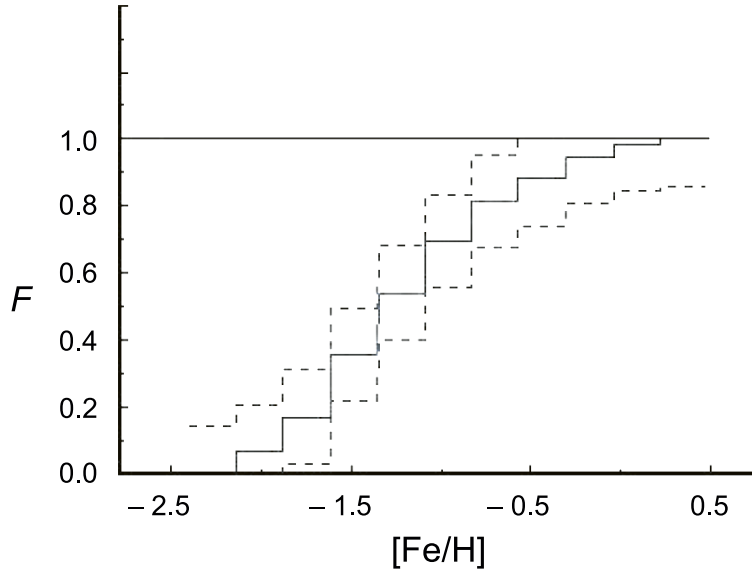

Рис. 3.2. Функция распределения металличности [Fe/H] для шаровых скоплений нашей Галактики

3.3. Оценивание плотности распределения

Основой для оценки плотности распределения является эмпирическая гистограмма — результат подсчетов попаданий измеренных значений в интервалы аргумента.

Одномерную гистограмму строят путем разбиения области, занимаемой выборкой $x_1, \ldots, x_n$ на m интервалов и подсчета числа попаданий элементов выборки в каждый k-й интервал (n_k). Общее количество интервалов m и их ширина $\Delta x = x_{k+1} - x_k$ выбираются из тех же соображений, что и при построении функции распределения. Далее речь пойдет *только о гистограммах с постоянной шириной интервалов*, к чему желательно стремиться для упрощения последующего анализа. В этом случае процесс построения гистограмм легко программируется. Иногда интервалы делают перекрывающимися, что является одним из способов сглаживания зависимостей.

От обычной гистограммы легко перейти к оценкам **вероятности попадания в данный интервал** по формуле

$$\breve{p}_k = \frac{n_k}{n}\,, \tag{3.26}$$

а также к оценкам **средней плотности вероятности** внутри каждого k-го интервала гистограммы

$$\breve{f}_k = \frac{\breve{p}_k}{\Delta x}\,. \tag{3.27}$$

Величины n_k, $\breve{p}_k$ и $\breve{f}_k$ являются *случайными*, то есть могут изменяться от выборки к выборке непредсказуемым образом.

Ту же методику построения гистограммы следует применять для многомерной случайной величины, только ширина интервала заменяется *объемом* m-мерного параллелепипеда.

Доверительная область для истинных значений средних по интервалам плотностей вероятности должна строиться в m-мерном пространстве, однако такую область невозможно представить наглядно, поэтому в практических исследованиях получила распространение *двумерная* доверительная область, обычно называемая **коридором ошибок**. При построении коридора ошибок, как правило, не переходят от n_k к $\breve{f}_k$, а строят этот коридор из доверительных интервалов для математических ожиданий величин n_k, то есть для

$$\hat{\mathrm{E}}(n_k) \equiv n\,\nu_k = n\,\breve{f}_k\,\Delta x\,. \tag{3.28}$$

Для построения k-го доверительного интервала можно воспользоваться нормальным распределением, поскольку случайная величина n_k, *независимо от распределения генеральной совокупности* $\vec{\mathbf{X}}$, имеет *биномиальное распределение* с математическим ожиданием

$$\nu_k = n\,p_k \tag{3.29}$$

и дисперсией

$$\hat{\mathrm{D}}(n_k) = n\,p_k\,(1-p_k) = \nu_k\left(1-\frac{\nu_k}{n}\right)\,, \tag{3.30}$$

но, в силу *предельной теоремы Муавра—Лапласа*, при большом объеме выборки *биномиальное распределение сходится к нормальному.*

Определим доверительную вероятность γ для гистограммы в целом как вероятность такого совместного события, как попадание истинной гистограммы одновременно во все доверительные интервалы, построенные с доверительной вероятностью γ' для отдельных столбцов гистограммы. Таким образом, можно записать

$$\gamma = \left(\gamma'\right)^m . \tag{3.31}$$

Если при построении доверительной полосы для гистограммы задать общую доверительную вероятность γ, то доверительная вероятность γ', для которой необходимо определить доверительные интервалы каждого столбца гистограммы, равна

$$\gamma' = \gamma^{\,1/m} . \tag{3.32}$$

При определении доверительных интервалов, как правило, используются значения доверительной вероятности γ, близкие к единице. Разлагая выражение (3.32), представленное в виде $(1-(1-\gamma))^{1/m}$, в степенной ряд по малому параметру $1-\gamma$ и ограничиваясь первым членом разложения, получим

$$\gamma' = 1 - \left(\frac{1-\gamma}{m}\right) . \tag{3.33}$$

В таком приближенном виде формула (3.32) приводится в [18].

Следуя общему правилу построения доверительных интервалов для математического ожидания и заменяя ν_k его оценкой n_k, получаем

$$n_k \pm u_{\gamma'} \sqrt{n_k \left(1 - \frac{n_k}{n}\right)} , \tag{3.34}$$

где $u_{\gamma'}$ — γ'-процентная точка нормального распределения, вычисленная для γ' по (3.32). На рис. 3.3 сплошной линией показана гистограмма распределения металличностей шаровых скоплений, а штриховые линии ограничивают коридор ошибок для 87 %-го доверительного интервала ($\gamma = 0.87$). Число измерений n, использованных для построения гистограммы, равно 135; число интервалов гистограммы $m = 14$. Доверительная вероятность для отдельного интервала гистограммы $\gamma' = 0.99$; значение γ'-процентной точки нормального распределения $u_{\gamma'} = 2.326$.

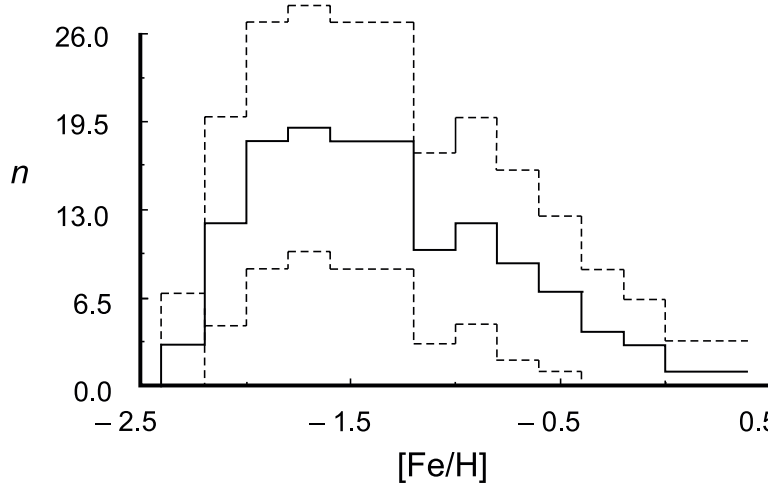

Рис. 3.3. Гистограмма распределения металличностей шаровых скоплений Галактики

Анализ выражения (3.34) показывает, что чем меньше число интервалов, для которого строится гистограмма, тем относительно у́же коридор ошибок. Однако укрупнение интервалов приводит к потере информации о тонкой структуре плотности распределения. Так, небольшой провал на гистограмме (рис. 3.3) около $[\mathrm{Fe/H}] \approx -1.1$ может означать, что распределение металличностей шаровых скоплений должно описываться суммой двух распределений с разными математическими ожиданиями. Укрупнение интервалов скроет эту возможность. При этом ясно, что провал на распределении меньше коридора ошибок и поэтому не является статистически значимым для принятого уровня значимости.

Рассмотренные оценки одномерной плотности вероятности дают выборочные значения средней по каждому из интервалов гистограммы плотности. Часто эти значения относят к серединам интервалов, считая их оценками плотности вероятности в точках x_j, что не совсем верно, так как значения $\breve{f}_j$, отнесенные к серединам интервалов, есть выборочные значения свертки двух функций — истинной плотности $f(x)$ и плотности равномерного распределения на интервале Δx. Такое расхождение между $\breve{f}_j$ и оценкой истинной плотности называют ***ошибкой интервала***.

Чтобы понять, как следует учитывать ошибку интервала, рассмотрим некоторый j-й интервал, схематически изображенный на рис. 3.4, где $f_j = \hat{\mathrm{E}}(\breve{f}_j)$ — математическое ожидание средней по данному интервалу плотности вероятности, а $f(x_j)$ — истинное значение плотности в центре интервала. Задача заключается в том, чтобы по оценкам $\breve{f}_j$ найти оценки значений $f(x_j)$. Из рисунка видно, что величина $f_j \Delta x$ равна

интегралу от $f(x)$ в пределах от $x_j - \Delta x/2$ до $x_j + \Delta x/2$. Разложив $f(x)$ в степенной ряд в окрестностях точки x_j и выполнив интегрирование в указанных пределах, получим

$$f_j = \sum_{k=0}^{\infty} \frac{f^{(2k)}(x_j)}{(2k+1)!} \left(\frac{\Delta x}{2}\right)^{2k}. \tag{3.35}$$

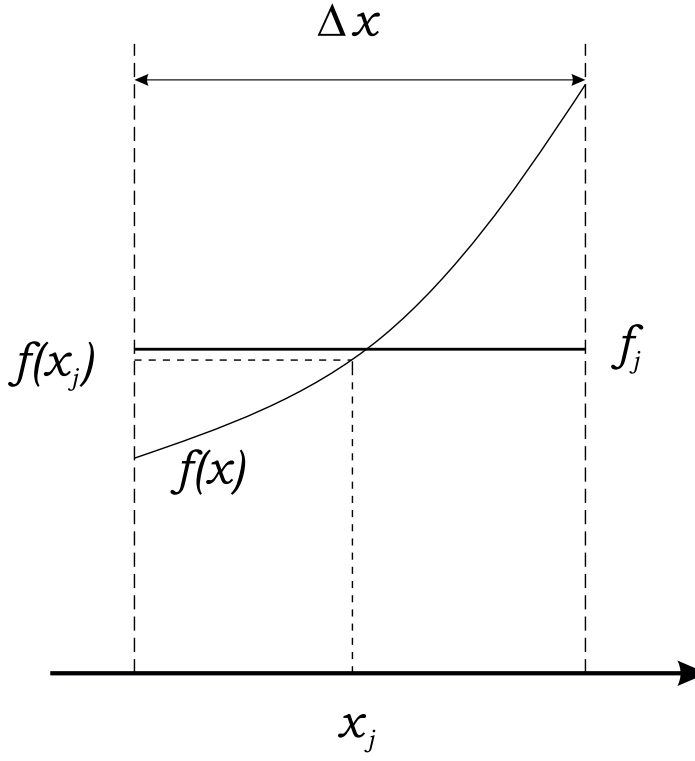

Рис. 3.4. Определение ошибки интервала для некоторого j-го интервала

Решая уравнение (3.35) относительно $f(x_j)$ методом последовательных приближений и заменяя f_j на $\breve{f}_j$, а производную f_j'' на $\Delta^2 \breve{f}_j/(\Delta x)^2$, где $\Delta^2 \breve{f}_j$ — вторая табличная разность, получим приближенное выражение

$$\breve{f}(x_j) \approx \breve{f}_j - \frac{\Delta^2 \breve{f}_j}{24}\,. \tag{3.36}$$

Можно использовать и упоминающиеся ниже формулы численного дифференцирования.

Иногда задачу об ошибке интервала решают несколько иначе. Например, П. Н. Холопов при исправлении кривых видимой звездной плотности в шаровых и рассеянных звездных скоплениях за ошибку интервала вводил поправки не в значения плотности, а в значения аргумента, что несколько менее удобно, так как обычно гистограммы распределений строят с использованием интервалов равной ширины для удобства дальнейшей обработки, а сдвиги аргумента нарушают это равенство.

Можно предложить принципиально иной способ избавления от ошибки интервала, *основанный на определении плотности вероятности как производной от функции распределения по ее аргументу*. Дело в том, что эмпирическая функция распределения (3.20) не подвержена влиянию ошибки интервала, поскольку сам способ ее построения дает значения этой функции не на интервалах, а в точках x_j. Поэтому достаточно численно продифференцировать $F(x_j)$ в точках x_j, чтобы получить оценки плотности вероятности $f(x_j)$. Для численного дифференцирования лучше всего применять специально разработанные для этого *сглаживающие выражения*, так как численное дифференцирование существенно увеличивает ошибки,

неизбежно присутствующие в функциях, получаемых из наблюдений. Для функций, полученных с постоянным шагом по аргументу, можно использовать выражение

$$f(x_j) \approx \frac{1}{10\,\Delta x}\,(-2F(x_{j-2}) - F(x_{j-1}) + \\ +F(x_{j+1}) + 2F(x_{j+2}))\ , \tag{3.37}$$

где $\Delta x = x_{j+1} - x_j = \text{const}$. Как все подобные формулы, выражение (3.37) не дает возможности вычислить значения $f(x_j)$ в крайних точках последовательности значений плотности распределения. Это связано с необходимостью использования информации о крайних точках для вычисления значений производной в следующих точках. Однако в данном случае мы можем, пользуясь свойствами функции распределения, доопределить последовательность, введя две новые вспомогательные точки слева, полагая в них $F(x_{-2}) = F(x_{-1}) = 0$, и две точки справа, полагая в них $F(x_{m+1}) = F(x_{m+2}) = 1$.

Следует отметить, что операция численного дифференцирования увеличивает случайную составляющую изменения функции (ведет к росту имеющихся в значениях функции ошибок), поэтому ее бессмысленно проводить при ощутимых разбросах точек у случайных функций.

3.4. Оценивание плотности распределения при наличии фона

В астрономии встречаются случаи, когда значения n_j на гистограмме искажены случайной поправкой. Наглядным примером являются рассеянные звездные скопления, для которых

невозможно построить ни одного эмпирического распределения, не искаженного примесью звезд галактического фона, которые внешне неотличимы от членов скопления. Исключение влияния таких звезд на результат исследования является важной задачей. Очевидно, что исключить влияние фона можно только статистически, оценив распределение изучаемой характеристики для объектов фона, так как имеющиеся собственные движения, с помощью которых можно выделить члены скопления, достаточно точны и многочисленны для очень небольшого количества объектов.

Рассмотрим в качестве примера оценку функции светимости звездного скопления по данным фотометрии звезд в поле скопления.

Для оценивания функции светимости членов скопления сначала строят две гистограммы распределений звезд по их видимым звездным величинам: одну для минимальной области, содержащей скопление, другую для области сравнения в ближайших окрестностях скопления. Обычно первая область ограничена окружностью с радиусом, определенным из звездных подсчетов, а вторая — кольцо вокруг скопления. Очевидно, что оценка числа звезд — членов скопления с лежащими в данном интервале видимыми звездными величинами есть

$$\breve{n}_j = n_j - n'_j \frac{a}{a'}\,, \tag{3.38}$$

где $\breve{n}_j$ — оценка числа членов скопления в данном интервале видимых звездных величин j; n_j и n'_j — такие же оценки для области скопления и фона соответственно; a и a' — площади области скопления и фоновой площадки. Для оценки точности

полученного результата вспомним, что при условии некоррелированности случайных величин n_j и n'_j дисперсия определяется суммой дисперсий слагаемых

$$\hat{\mathrm{D}}(\breve{n}_j) = \hat{\mathrm{D}}(n_j) + \left(\frac{a}{a'}\right)^2 \hat{\mathrm{D}}(n'_j)\,, \tag{3.39}$$

где $\hat{\mathrm{D}}$ — оператор дисперсии.

Покажем теперь, что *при равномерном распределении* плотности звезд фона дисперсия числа звезд в некоторой площадке, связанная со случайными флуктуациями фона, пропорциональна площади этой площадки. Разобьем фоновую площадку с площадью a на t площадок одинаковой площади Δa. Пусть n — число звезд во всей области, а n_k — в k-й площадке ($k = 1,\ 2,\ \ldots,\ t$). Обозначим через $\sigma^2_{\Delta a}$ дисперсию случайной величины n_k. Поскольку $n = n_1 + n_2 + \ldots + n_t$, то дисперсия числа звезд во всей области, при условии некоррелированности случайных величин n_k, есть

$$\sigma^2_a = t\,\sigma^2_{\Delta a} = \sigma^2_{\Delta a}\,\frac{a}{\Delta a}\,, \tag{3.40}$$

$$\frac{\sigma^2_a}{a} = \frac{\sigma^2_{\Delta a}}{\Delta a} \equiv \sigma^2_0\,, \tag{3.41}$$

где σ^2_0 — дисперсия числа звезд на единицу площади. Отсюда следует, что

$$\sigma^2_a = a\,\sigma^2_0 \tag{3.42}$$

для любой области площади a.

Теперь формулу (3.39) для оценки дисперсии «снизу» можно представить в окончательном виде:

$$\hat{\mathrm{D}}(\breve{n}_j) \simeq s^2_{0,\,j}\,a + s^2_{0,\,j}\left(\frac{a}{a'}\right)^2 a' = s^2_{0,\,j}\,a\left(1+\frac{a}{a'}\right), \qquad (3.43)$$

где $s^2_{0,\,j}$ — оценка дисперсии числа звезд фона на единицу площади для j-го интервала звездных величин. Для получения этой оценки можно разбить область сравнения (фона) на t равных по площади участков и вычислить дисперсию $s^2_{0,\,j}$ для каждого интервала звездных величин по обычной формуле для выборочной дисперсии с учетом того, что площадь каждого участка есть $a'/\,t$. В нашем случае указанная формула принимает вид

$$s^2_{0,\,j} = \frac{t}{(t-1)\,a'}\sum_{k=1}^{t}\left(n_{jk} - \frac{n'_j}{t}\right)^2, \qquad (3.44)$$

где $n'_j/\,t$ представляет собой *среднее число звезд фона* в участке с площадью $a'/\,t$; индекс k является номером участка. Как можно видеть из формул (3.43) и (3.44), для повышения точности оценивания функции светимости звездного скопления площадь области сравнения должна быть по возможности большей, а площадь области скопления должна быть минимального размера, но не менее размеров скопления.

Задача, подобная изложенной, возникает при исследовании диаграмм *показатель цвета — звездная величина* шаровых скоплений, когда звезды фона убираются с диаграммы статистическим методом. Для этого строится путем подсчетов двумерное распределение звезд фона на диаграмме *показатель цвета—звездная величина* с помощью фотометрии звезд в лежащих недалеко от скоплений площадок сравнения, и эта плот-

ность звезд фона вычитается из плотности звезд скопления с учетом отношения площадей площадки скопления и площадки фона. Единственное различие с подробно рассмотренным случаем оценивания функций светимости заключается в том, что при «чистке» фотометрических диаграмм звездная плотность не является двумерным равномерным распределением. К сожалению, подобная методика малоприменима в случае рассеянных скоплений из-за меньшего числа членов скоплений и больших флуктуаций свойств распределений фоновых звезд в площадках, расположенных вблизи плоскости Галактики, но для отдельных богатых звездами рассеянных скоплений ее применение оправдано.

Подобная задача возникает и при определении пространственной плотности звезд в звездных скоплениях. Она будет подробно рассмотрена в одной из следующих глав.

4. ИСПРАВЛЕНИЕ НАБЛЮДАЕМЫХ РАСПРЕДЕЛЕНИЙ ЗА СЛУЧАЙНЫЕ ОШИБКИ

4.1. Влияние случайных ошибок на выборочные распределения

При рассмотрении многих вопросов звездной статистики мы часто пользуемся множеством измерений какой-либо величины, и нас интересует *плотность распределения этой величины, освобожденная от ошибок наблюдений*. Поэтому наблюдаемое распределение необходимо каким-либо образом «исправить» за ошибки наблюдений.

Пусть ошибки наблюдений распределены по нормальному закону $N\left(0,\, \sigma^2\right)$ с известной дисперсией. Тогда вероятность $P(\epsilon)$ ошибки, заключенной в интервале $[\,\epsilon,\, \epsilon + d\epsilon\,]$, выражается формулой

$$P(\epsilon)\, d\epsilon = \frac{h}{\sqrt{\pi}}\, \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon\,, \tag{4.1}$$

причем здесь для удобства записи использована величина $h = \dfrac{1}{\sigma\sqrt{2}}$, называемая ***мерой точности***. Пусть функ-

ция $u(x)$ будет действительным («исправленным») распределением величины x, а функция $v(x)$ — наблюдаемым распределением, искаженным ошибками наблюдений. Полученное из наблюдений значение x включает в себя ошибку ϵ, следовательно, точное значение t этой величины есть $t = x - \epsilon$.

Число объектов, точные величины которых заключены в пределах $[\,t,\ t + dt\,]$, а ошибка — в пределах $[\,\epsilon,\ \epsilon + d\epsilon\,]$, равно

$$u(x)\,dx\,d\epsilon = \frac{h}{\sqrt{\pi}}\,u(t)\,\exp\left(-h^2\,\epsilon^2\right)\,dt\,d\epsilon\,. \tag{4.2}$$

Проведем замену переменных t и ϵ на x и ϵ, при этом очевидно $dx\,d\epsilon = dt\,d\epsilon$. Следовательно, относительное число звезд с наблюденной величиной x в пределах от x до $x+dx$ и ошибкой в пределах от ϵ до $\epsilon + d\epsilon$ определяется формулой

$$v(x)\,dx\,d\epsilon = \frac{h}{\sqrt{\pi}}\,u(x-\epsilon)\,\exp\left(-h^2\,\epsilon^2\right)\,dx\,d\epsilon\,. \tag{4.3}$$

Интегрируя последнее выражение по всем ϵ от $-\infty$ до $+\infty$, получим общее число звезд с характеристикой x, заключенной в пределах от x до $x + dx$:

$$v(x)\,dx = \frac{h}{\sqrt{\pi}}\,dx\int\limits_{-\infty}^{\infty} u(x-\epsilon)\,\exp\left(-h^2\,\epsilon^2\right)\,d\epsilon\,. \tag{4.4}$$

Таким образом, *наблюдаемое распределение действительно является сверткой истинного распределения с распределением ошибок*, в данном случае распределенными по нормальному закону. Решив это уравнение, можно определить истинное распределение, используя информацию о распределении ошибок.

Из приведенного вывода уравнения свертки (4.4) видно, что распределение ошибок может быть любым, в том числе таким, где мера точности (дисперсия) является функцией аргумента, что часто встречается в приложениях. Так, обычно расстояния до более далеких звезд и звездных скоплений определяется в среднем с большей ошибкой.

Чтобы понять, что происходит с плотностью распределения под влиянием случайных ошибок или, говоря иным языком, как действует на функцию операция свертки, рассмотрим рис. 4.1. На рисунке изображена часть гистограммы выборочного распределения, содержащая три столбца. Под действием ошибок определенная доля точек из одного интервала (столбца гистограммы) переходит в соседний. Но из более высоких столбцов в менее высокие переходит больше точек, чем в обратном направлении. Таким образом, более высокие столбцы становятся ниже, а менее высокие — выше, *распределение расплывается, становится шире. Это соответствует увеличению дисперсии распределения согласно величине дисперсии ошибок.* При этом возможные провалы или выступы в реальном распределении замываются, мелкие детали пропадают. Следствием этих эффектов могут быть значительные искажения информации, получаемой из обработки наблюденных данных. Например, длины наблюдаемых векторов собственных движений космических объектов в среднем всегда больше, чем реальные. То же можно сказать о величинах пространственных скоростей.

Рассмотрим еще один пример. Пусть мы решаем задачу звездной кинематики — определяем *постоянные Оорта*. Для решения задачи мы имеем выборку звездных объектов с извест-

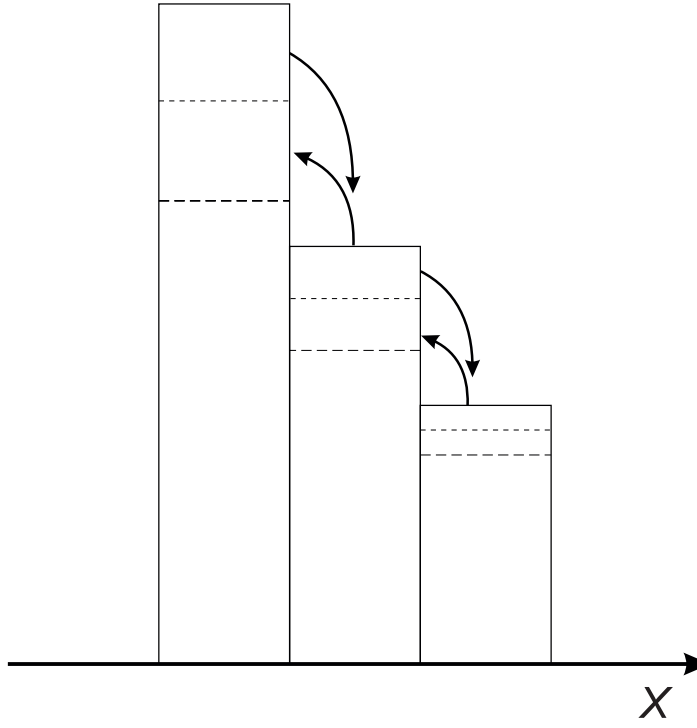

Рис. 4.1. Влияние ошибок измерения на гистограмму выборочного распределения

ными расстояниями. Но расстояния определяются с заметной ошибкой. В результате распределение расстояний объектов выборки от Солнца расплывается, как было показано выше, и мы используем преувеличенные в среднем расстояния до объектов. Очевидно, что исправлять наблюдаемые (выборочные) распределения за случайные ошибки совершенно необходимо. Перейдем к методам решения этой задачи.

4.2. Приближенный метод исправления распределений

Название этого раздела несколько условно, так как все методы статистики следует считать приближенными, ибо решения задач статистики получаются как решения, с определенной вероятностью лежащие внутри определенного доверительного интервала. В данном случае под приближенным решением мы понимаем случай, когда при разложении функций в ряды для решения оставляют определенное количество первых членов ряда.

Рассмотрим выражение (4.4), где левая и правая части равенства разделены на dx. Предположим, что ошибки наблюдений малы. Разложим в ряд по малому параметру ϵ стоящую под знаком интеграла функцию $u(x)$. Получим

$$v(x) = \frac{h}{\sqrt{\pi}} \int\limits_{-\infty}^{\infty} \left(u(x) - \epsilon\, u_1(x) + \right. \\ \left. + \frac{\epsilon^2}{2!}\, u_2(x) - \ldots \right) \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon\,, \tag{4.5}$$

где введено обозначение $u_n \equiv \dfrac{d^n u}{dx^n}$. Принимая во внимание, что для целых p

$$\int\limits_{-\infty}^{\infty} \epsilon^{2p+1}\, \exp\left(-h^2 \epsilon^2\right) d\epsilon = 0\,, \tag{4.6}$$

получим, что выражение (4.5) принимает вид

$$v(x) = \frac{h}{\sqrt{\pi}} \int\limits_{-\infty}^{\infty} \left(u(x) + \frac{\epsilon^2}{2!}\, u_2(x) + \right. \\ \left. + \frac{\epsilon^4}{4!}\, u_4(x) + \ldots \right) \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon\,. \tag{4.7}$$

Введем обозначение

$$A \equiv \int\limits_{-\infty}^{\infty} \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon = \frac{\sqrt{\pi}}{h}\,. \tag{4.8}$$

Так как A является *равномерно сходящимся интегралом*, то можно провести дифференцирование по h под знаком интеграла и записать

$$\frac{dA}{dh} = -2h \int\limits_{-\infty}^{\infty} \epsilon^2\, \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon = -\frac{\sqrt{\pi}}{h^2}\,. \tag{4.9}$$

Из чего следует

$$\int\limits_{-\infty}^{\infty} \epsilon^2\, \exp\left(-h^2\, \epsilon^2\right)\, d\epsilon = \frac{\sqrt{\pi}}{2h^3}\,. \tag{4.10}$$

Продолжая дифференцирование, придем наконец к общей формуле

$$\int\limits_{-\infty}^{\infty} \epsilon^{2p}\, \exp(-h^2\, \epsilon^2)\, d\epsilon = \frac{\sqrt{\pi}}{h}\, \frac{(2p)!}{p!\, (4h^2)^p}\,. \tag{4.11}$$

С помощью этого выражения равенство (4.7) после раскрытия скобок и интегрирования переписывается в виде

$$v(x) = u(x) + \frac{1}{4\,h^2}\,u_2(x) + \\ + \frac{1}{32\,h^4}\,u_4(x) + \ldots + \frac{1}{p!\;(4h^2)^p}\,u_{2p}(x)\,, \tag{4.12}$$

причем *нет смысла оставлять в отрезке ряда больше членов, чем содержится интервалов аргумента в гистограмме рассматриваемого распределения*. При достаточно больших значениях h (малых значениях дисперсии) мы можем найти $u(x)$ из (4.12) методом последовательных приближений.

I приближение:

$$u(x) = v(x)\,; \tag{4.13}$$

II приближение:

$$u^{(1)}(x) = v(x) - \frac{1}{4h^2}\,v_2(x)\,; \tag{4.14}$$

III приближение:

$$u^{(2)}(x) = u^{(1)}(x) - \frac{1}{4h^2}\,u_2^{(1)}(x) - \frac{1}{32h^4}\,u_4^{(1)}(x) \tag{4.15}$$

и т. д. Здесь нижние индексы при функциях вновь обозначают соответствующие производные. Процесс выписывания следующих приближений можно продолжать, хотя использование производных более высокого порядка, которые приходится для выборочных произведений брать численно, вряд ли имеет смысл. Производные вычисляются либо с помощью выражений типа (3.37), либо для этого используются табличные разности

соответствующих порядков. Так, с помощью разностей до четвертого порядка из (4.15) получим

$$u(x) = v(x) - \frac{12\Delta_2(x) - \Delta_4(x)}{48h^2\,a^2} - \frac{\Delta_4}{32h^4\,a^4}\,, \qquad (4.16)$$

где a — ширина шага таблицы (интервал гистограммы выборочного распределения). О более точных (сглаживающих) методах численного дифференцирования будет сказано в одной из следующих глав.

Некоторые исследователи кинематики Галактики, такие как М. Фист, М. Шаттлворт, Л. Балона и Д. Крэмптон, использовали для исправления распределений расстояний звезд выборок от Солнца полученные на основе изложенного здесь метода исправляющие множители. На эти множители, меньшие единицы и зависящие от расстояния, умножались расстояния звезд выборки, чтобы компенсировать в среднем рост расстояний от Солнца, вызываемый уширением распределения за счет влияния случайных ошибок. Такой метод учета влияния случайных ошибок позволял получить более надежные и несмещенные оценки параметров кинематических моделей Галактики. Однако метод исправляющих множителей не совсем надежен, так как исправляющие множители, выводимые при определенных допущениях, оказывались не зависящими от формы распределения расстояний, что не дает возможности адекватно исправить ситуацию.

4.3. «Точные» методы исправления выборочных распределений

Первым «точным» методом исправления при не очень больших ошибках наблюдений можно считать метод, связанный с *параметризацией наблюдаемого распределения*. В первом параграфе данной главы мы рассмотрели выбор кривой Пирсона, приближающей выборочную плотность распределения. Пусть такая кривая выбрана этим или иным способом. *Предположим, что ошибки наблюдений настолько малы, что и истинное распределение принадлежит к тому же типу распределения, то есть ошибки влияют на выборочное распределение так, что не меняют его пирсоновского типа.* Тогда подбором параметров выбранной кривой Пирсона, выражение которой подставлено в уравнение (4.4) в качестве истинного распределения $u(x)$, можно добиться приближенного равенства левой и правой частей (4.4) и использовать в качестве «исправленного» распределение с новыми параметрами.

Следующим методом может служить *непосредственное решение уравнения свертки* (1.36) или, в случае нормального распределения ошибок, уравнения (4.4) путем использования свойств преобразования Фурье.

Определим **прямое преобразование Фурье** как

$$f(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) \exp(-i\omega x)\, dx \qquad (4.17)$$

и **обратное преобразование Фурье** как

$$f(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} f(\omega) \exp(i\omega x)\, d\omega\,. \tag{4.18}$$

Обозначим $\hat{\mathrm{F}}\,[\,f\,]$ *преобразование Фурье* функции f (не путать с функцией распределения), а знаком $*$ обозначим **операцию свертки**. Операцию свертки будем понимать как

$$u(x) * g(x) = \int\limits_{-\infty}^{+\infty} f(x-\epsilon) g(\epsilon)\, d\epsilon\,. \tag{4.19}$$

В таком случае уравнение (1.36) запишется в виде

$$v(x) = u(x) * g(x)\,, \tag{4.20}$$

где $g(x)$ — распределение ошибок. Используем свойство преобразования Фурье свертки двух функций (*теорема о свертке*).

Теорема 4.1. *Фурье-преобразование свертки двух функций есть нормированное произведение Фурье-преобразований этих функций.*

$$\hat{\mathrm{F}}\,[\,f * g\,] = \sqrt{2\pi}\,\hat{\mathrm{F}}\,[\,f\,]\;\hat{\mathrm{F}}\,[\,g\,]\;. \tag{4.21}$$

Если известно распределение ошибок, то, применяя свойство (4.21) к уравнению (4.20), получаем

$$\hat{\mathrm{F}}\,[\,u(x) * g(x)\,] = \sqrt{2\pi}\,\hat{\mathrm{F}}\,[\,u(x)\,]\;\hat{\mathrm{F}}\,[\,g(x)\,]\;, \tag{4.22}$$

откуда имеем

$$\hat{\mathrm{F}}\,[\,u\,] = \frac{\hat{\mathrm{F}}\,[\,u * g\,]}{\sqrt{2\pi}\,\hat{\mathrm{F}}\,[\,g\,]}\,. \tag{4.23}$$

Но, в силу уравнения (4.20),

$$\hat{\mathrm{F}}\,[\,v(x)\,] = \hat{\mathrm{F}}\,[\,u * g\,]\;, \tag{4.24}$$

поэтому окончательно

$$\hat{\mathrm{F}}\,[\,u(x)\,] = \frac{\hat{\mathrm{F}}\,[\,v(x)\,]}{\sqrt{2\pi}\,\hat{\mathrm{F}}\,[\,g(x)\,]}\,. \tag{4.25}$$

Теперь для получения распределения, исправленного за влияние случайных ошибок, достаточно применить к (4.25) обратное Фурье-преобразование.

Так как выборочная плотность распределения обычно представляется *в дискретном виде* — в виде гистограммы, приведем формулы для дискретного прямого и обратного преобразований Фурье.

Определение 4.1. *Пусть на N значениях аргумента $x_k = x_0 + k\Delta x$, где Δx — шаг гистограммы, задана функция (значения гистограммы) своими значениями*

$$f_k = f(x_k), \quad k = 0,\ 1,\ \ldots,\ N-1. \tag{4.26}$$

Пусть $\Delta x = T/(N-1)$, где T — период функции (размах гистограммы).

Фурье-преобразование вектора данных $\vec{\mathbf{F}}^T = (f_0,\ \ldots,\ f_{N-1})$ есть вектор $\vec{\boldsymbol{\phi}} = \hat{\Phi}\,\vec{\mathbf{F}}$, где $\hat{\Phi}$ — матрица размера $N \times N$ с эле-

ментами

$$\exp\left(-2\pi\, i\, \omega_m\, x_k\right)\,, \tag{4.27}$$

где i *— мнимая единица;* $\omega_m = m\,\Delta\omega$*;* $m = 0,\ \ldots,\ N-1$*;* $\Delta\omega = 1/T$.

Компоненты вектора $\vec{\phi}$ аналогичны коэффициентам Фурье в обычных тригонометрических разложениях. Обратное преобразование Фурье получается из прямого заменой в (4.27) знака в показателе степени экспоненты с минуса на плюс.

Отметим, что прямое дискретное преобразование Фурье требует N^2 арифметических операций. Поэтому в настоящее время обычно применяют так называемое *быстрое преобразование Фурье* (в частности, алгоритм Кули—Тьюки [19, 20]), для чего разработано много разнообразных программ. Но при не очень большом числе интервалов гистограммы можно использовать непосредственно дискретное преобразование Фурье.

В заключение заметим, что в случае нормального распределения ошибок использование преобразования Фурье еще более упрощается, так как преобразование Фурье нормальной плотности распределения с нулевым средним и дисперсией σ^2 также есть нормальное распределение, но с дисперсией $1/\sigma^2$.

5. ДИСПЕРСИОННЫЙ АНАЛИЗ

Определение 5.1. *Дисперсионный анализ — это статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента.*

Первоначально дисперсионный анализ был предложен английским генетиком и статистиком Р. Фишером для обработки результатов агрономических опытов по выявлению условий, при которых испытываемый сорт сельскохозяйственной культуры дает максимальный урожай. В астрономической практике напрямую дисперсионный анализ применяется редко, однако элементы дисперсионного анализа входят во многие более часто используемые методы статистического анализа данных, поэтому мы начинаем обзор таких методов именно с дисперсионного анализа.

Рассмотрим статистическую гипотезу

$$H_0 : \ \xi_1 = \xi_2 = \ldots = \xi_m$$

о равенстве математических ожиданий для m одномерных генеральных совокупностей для случайных величин $X_1, \ \ldots, \ X_m$. Можно рассмотреть также другой вариант задачи, когда ге-

неральная совокупность одна, но имеется подозрение, что ее математическое ожидание меняется под влиянием некоторых факторов, количественных или качественных.

Дисперсионный анализ можно рассматривать как метод проверки указанных гипотез по исследованию влияния прежде всего качественных факторов на результаты экспериментов, так как при количественных факторах уместнее использовать ***регрессионный анализ***, изложенный в следующих главах.

Примером качественных факторов может служить класс объекта, то есть его принадлежность к той или иной совокупности. Наиболее очевидное поле непосредственного применение дисперсионного анализа в звездной астрономии — случай объединения величин из разных каталогов, содержащих данные наблюдений, в более общие компилятивные каталоги, когда объединяются данные, полученные на разных инструментах разных обсерваторий в разное время. Дисперсионный анализ в этом случае сигнализирует о возможной неоднородности получающегося каталога и необходимости поиска и исключения систематических ошибок в отдельных рядах данных.

Проверку гипотезы $H_0: \xi_1 = \xi_2 = \ldots = \xi_m$ можно было бы провести путем попарного сравнения всех возможных пар случайных величин $X_1, \ldots, X_m$, однако такая процедура очень трудоемка.

Идея метода дисперсионного анализа основана на сведении неслучайных различий между $\xi_1, \xi_2, \ldots, \xi_m$ к случайным с последующим анализом выборочных дисперсий.

Рассмотрим этот метод на примере **однофакторного дисперсионного анализа**, когда исследуемые *генеральные совокупности отличаются друг от друга только одним каче-*

ственным признаком или если генеральная совокупность одна, а на ее математическое ожидание влияет *только один внешний фактор* t, принимающий дискретные значения $t_1,\ t_2,\ \ldots,\ t_m$. В этом, последнем случае гипотезу следует записать в форме

$$H_0:\ \xi(t_1) = \xi(t_2) = \ldots = \xi(t_m)\,. \tag{5.1}$$

Итак, пусть имеем m нормально распределенных случайных величин $X_1,\ \ldots,\ X_m$ с одинаковыми генеральными дисперсиями σ_0^2. По m выборкам с объемами $n_1,\ \ldots,\ n_m$ найдем выборочные оценки математических ожиданий $\langle x\rangle_k$ и дисперсий $s^2_{0,\,k}$ $(k = 1,\ \ldots,\ m)$. При справедливости гипотезы $H_0:\ \xi_1 = \xi_2 = \ldots = \xi_m$ должно выполняться равенство $\langle x\rangle_1 = \langle x\rangle_2 = \ldots = \langle x\rangle_m$, а выборки можно объединить в одну и найти общую оценку дисперсии s^2_{all}, которая должна *незначимо* отличаться от каждой из оценок $s^2_{0,\,k}$ или от их среднего взвешенного значения s_0^2. В противном случае различие указанных дисперсий окажется значимым, поскольку к случайному рассеянию элементов объединенной выборки добавится разброс, связанный с различием математических ожиданий $\xi_1,\ \xi_2,\ \ldots,\ \xi_m$.

Исходные данные для дисперсионного анализа обычно располагают в **таблице данных** (табл. 5.1).

Средние значения по столбцам таблицы $\langle x\rangle_k$, показанные в последней строке, а также общее среднее $\langle x\rangle$ вычисляются

Таблица 5.1

Таблица данных для дисперсионного анализа

	k				
i	1	2	3	...	m
1	x_{11}	x_{12}	x_{13}	...	x_{1m}
2	x_{21}	x_{22}	x_{23}	...	x_{2m}
...	...	...	...	...	...
n_2	...	$x_{n_2 2}$	...	...	...
...	...		...	...	...
$n_1 = n_3$	$x_{n_1 1}$		$x_{n_3 3}$	...	...
...				...	...
n_m				...	$x_{n_m m}$
Средние	$\langle x \rangle_1$	$\langle x \rangle_2$	$\langle x \rangle_3$	...	$\langle x \rangle_m$

по обычным формулам:

$$\langle x \rangle_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ik} \,, \tag{5.2}$$

$$\langle x \rangle = \frac{1}{N} \sum_{k=1}^{m} \sum_{i=1}^{n_k} x_{ik} \,, \tag{5.3}$$

где $N = n_1 + n_2 + \ldots + n_m$.

Рассмотрим один из элементов нашей выборки, например, x_{ik}. Для этого элемента можно записать очевидное соотношение

$$x_{ik} - \langle x \rangle = (\langle x \rangle_k - \langle x \rangle) + (x_{ik} - \langle x \rangle_k) \,, \tag{5.4}$$

то есть *общее отклонение* элемента x_{ik} от общего среднего равно сумме *отклонения столбца* $(\langle x \rangle_k - \langle x \rangle)$ и *остаточного от-*

клонения в столбце $(x_{ik} - \langle x \rangle_k)$. Если соотношение (5.4) возвести в квадрат и просуммировать по всем индексам, получим так называемое **дисперсионное соотношение**

$$\underbrace{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\left(x_{ik} - \langle x \rangle\right)^2}_{SS_{\text{общ}}\text{ — общее}} = \underbrace{\sum_{k=1}^{m} n_k \left(\langle x \rangle_k - \langle x \rangle\right)^2}_{SS_{\text{ст}}\text{ — по столбцам}} + \underbrace{\sum_{k=1}^{m}\sum_{i=1}^{n_k}\left(x_{ik} - \langle x \rangle_k\right)^2}_{SS_0\text{ — остаточное}}, \tag{5.5}$$

и, обозначив, как принято в математической статистике, суммы квадратов буквами SS (от англ. sum of squares — сумма квадратов), запишем

$$SS_{\text{общ}} = SS_{\text{ст}} + SS_0\,. \tag{5.6}$$

Здесь величины SS соответствуют суммам в выражении (5.5).

Соотношения (5.4) и (5.5) удобно представить в виде таблицы дисперсионного анализа (табл. 5.2).

Из соотношения (5.5) видно, что при справедливости гипотезы о равенстве математических ожиданий случайных величин $X_1,\ X_2,\ \ldots,\ X_m$ сумма квадратов по столбцам должна быть небольшой, незначимой по сравнению с остаточной суммой квадратов, поэтому в качестве критерия проверки гипотезы можно взять отношение дисперсий, соответствующих этим суммам квадратов

$$F = \frac{s^2_{\text{ст}}}{s^2_0} = \frac{SS_{\text{ст}}}{(m-1)}\,\frac{(N-m)}{SS_0}\,. \tag{5.7}$$

Таблица 5.2

Таблица дисперсионного анализа

Отклонения	Сумма квадратов	Степени свободы
По столбцам	$SS_{\text{ст}} = \sum\limits_{k=1}^{m} n_k \left(\langle x\rangle_k - \langle x\rangle\right)^2$	$m-1$
Остаточное	$SS_0 = \sum\limits_{k=1}^{m} \sum\limits_{i=1}^{n_k} \left(x_{ik} - \langle x\rangle_k\right)^2$	$N-m$
Общее	$SS_{\text{общ}} = \sum\limits_{k=1}^{m} \sum\limits_{i=1}^{n_k} \left(x_{ik} - \langle x\rangle\right)^2$	$N-1$

В случае нормально распределенных случайных величин $X_1,\ X_2,\ \ldots,\ X_m$ статистика F распределена *по закону Фишера* с $m-1$ и $N-m$ степенями свободы, а критическая область при заданном уровне значимости определяется неравенством

$$F > F_\alpha(m-1,\ N-m)\,, \tag{5.8}$$

где $F_\alpha(m-1,\ N-m)$ — α-процентная точка распределения Фишера; $\alpha = 1-\gamma$; γ — выбранная доверительная вероятность. Выполнение этого неравенства означает, что по крайней мере одно из математических ожиданий значимо отличается от остальных. К сожалению, данный критерий не позволяет выявить, какие именно ξ_k оказались значимо отличными от остальных, однако существуют приемы, например *метод Шеффе* [21], также называемый *методом линейных контрастов*, в которых этот недостаток устранен.

В отличие от только что рассмотренного однофакторного **многофакторный дисперсионный анализ** рассматривает одномерную случайную величину X, на которую влияют m фак-

торов, каждый из которых может принимать l_k дискретных значений — **уровней**, а задача заключается в том, чтобы *статистически выявить значимые факторы.* Для решения задачи необходимо взять по одной выборке для каждой комбинации уровней факторов, то есть всего $l_1 \cdot l_2 \cdot \ldots \cdot l_m$ выборок. Такой способ проведения эксперимента, при котором перебираются все комбинации уровней факторов, называют **полным факторным экспериментом**. Число выборок можно резко сократить, прибегнув к так называемому **дробному эксперименту**, когда рассматриваются не все сочетания уровней факторов.

Отметим, что для дисперсионного анализа часто используют аббревиатуру ANOVA (от англ. analysis of variance), она встречается также в пакетах программ для статистической обработки данных.

6. РЕГРЕССИОННЫЙ АНАЛИЗ

6.1. Математические модели регрессии

Регрессионный анализ является наиболее часто используемым методом статистического анализа данных. Он нередко отождествляется с методом наименьших квадратов (МНК), что не совсем верно, поскольку МНК как метод получения наилучших в некотором смысле решений используется и в вычислительной математике, например, при аппроксимации функций. С другой стороны, в регрессионном анализе можно использовать не только МНК, но и другие методы получения оценок: метод минимакса, метод максимального правдоподобия, а также применить минимизацию сумм иных степеней уклонений. Метод наименьших квадратов имеет преимущество в том смысле, что дает наиболее простой метод отыскания параметров модели, выбранной нами для описания исследуемого явления при хорошем качестве получаемых оценок параметров.

Определение 6.1. *В* ***регрессионном анализе*** *рассматривается одномерная случайная величина, называемая* ***откликом*** *и играющая роль «зависимой» переменной (в смысле функциональной, а не статистической зависимости). Дисперсия этой величины в случае равноточных измерений по-*

*стоянна, а математическое ожидание меняется под воздействием одной или нескольких случайных величин — **факторов** («независимых» переменных). Принципиальным отличием регрессионного анализа от дисперсионного является то, что, во-первых, факторы здесь **всегда количественные**, и, во-вторых, задача регрессионного анализа заключается не только и не столько в выявлении значимости влияния факторов на отклик (обычно существование этой зависимости известно заранее), а в получении математической модели такого влияния и в оценивании неизвестных параметров модели с последующим анализом точности результатов.*

Определение 6.2. *Под **математической моделью** в регрессионном анализе понимают некоторую функцию от факторов, более или менее адекватно аппроксимирующую истинную (генеральную) поверхность регрессии* (1.18).

Вид этой функции, как правило, неизвестен, поэтому мы и говорим о модели. Целью построения модели обычно является ее дальнейшее использование в качестве эмпирической формулы для предсказания среднего значения отклика при заданных значениях факторов. Часто значения неизвестных параметров модели не имеют физического смысла. Но иногда нас интересуют именно значения параметров, а не эмпирическая формула, особенно в тех случаях, когда вид модели обусловлен теоретическими соображениями. В качестве примера здесь можно упомянуть об определении постоянных Оорта при изучении кинематики диска Галактики или построение зависимостей период—светимость для разных типов переменных звезд.

В первом случае вид связи лучевой скорости с расстоянием до объекта и положением в координатах l и b задан, а коэффициенты модели имеют определенный физический смысл. Во втором случае функциональное выражение дается теорией. Еще один пример: в звездной астрономии широко используется экспоненциальная модель

$$D = D_0 \exp\left(-\frac{|z|}{\beta}\right) \tag{6.1}$$

для аппроксимации пространственного распределения галактических объектов по z-координате. В этой модели параметры D_0 и β имеют вполне определенный физический смысл, их оценивание по выборке звезд или других объектов становится целью регрессионного анализа.

В качестве математических моделей в регрессионном анализе чаще всего используются *максимально простые модели*: полиномы, экспоненты, отрезки тригонометрических рядов и т. д. [5]. Оценка параметров модели оказывается особенно простой при линейной по параметрам модели, то есть сама модель может быть и нелинейной в отношении факторов (которые иногда называют независимыми переменными, что, как будет видно, вносит путаницу в описание моделей), но неизвестные параметры должны входить в выражение для модели линейно. Например, широко используемая в фотографической астрометрии полиномиальная модель второго порядка

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \tag{6.2}$$

является линейной по параметрам $\beta_0, \ldots, \beta_5$.

На примере модели (6.2) отметим разницу между понятиями «фактор» и «независимая переменная». В (6.2) мы имеем две независимые переменные, x_1 и x_2, а факторами являются $x_1,\ x_2,\ x_1^2,\ x_2^2,\ x_1x_2$, так что модель (6.2) является пятифакторной.

Отметим, что некоторые нелинейные модели можно свести к линейным. Так, модель (6.1) превращается в линейную после логарифмирования левой и правой частей равенства, при этом вместо вектора параметров $(D_0,\ \beta)$ мы будем искать вектор параметров $(\lg(D_0),\ 1/\beta)$ с возможностью легкого обратного перехода. Тем не менее следует помнить, что при таком преобразовании изменяется закон распределения ошибок измерения величины D, что может привести к смещению оценки. *Модели, сводимые к линейным путем каких-либо преобразований параметров, часто называют внутренне линейными.*

Определение 6.3. *Линейная по параметрам модель связи одномерной случайной величины Y с изменяющимися факторами $x_1,\ x_2,\ \ldots,\ x_m$ в общем случае имеет вид*

$$Y = \beta_0\, x_0 + \beta_1\, x_1 + \ldots + \beta_m\, x_m + \epsilon = \vec{\mathbf{X}}^T\, \vec{\boldsymbol{\beta}} + \epsilon\,, \qquad (6.3)$$

где $\vec{\boldsymbol{\beta}}^{\,T} = (\beta_0,\ \beta_1,\ \ldots,\ \beta_m)$ — вектор неизвестных параметров; $\vec{\mathbf{X}}^T = (x_0,\ x_1,\ \ldots,\ x_m)$ — вектор независимых переменных факторов, причем фиктивная переменная x_0, введенная для общности обозначений, всегда равна единице; ϵ — случайная величина с нулевым математическим ожиданием и дисперсией σ_0^2.

Например, для кинематической модели, описывающей вклад движения Солнца в пространстве в лучевые скорости

звезд окрестностей Солнца, имеем

$$V_{\mathrm{R}} = u\,\cos l\cos b + v\,\sin l\cos b + w\,\sin b + \epsilon\,. \qquad (6.4)$$

Здесь в модели содержатся три фактора, являющихся функциями от двух переменных l и b. Определяемыми параметрами являются u, v, w.

Отметим, что на месте факторов могут стоять некоторые не содержащие неизвестных параметров функции от независимых переменных.

6.2. Оценивание параметров линейной регрессионной модели

Пусть выбрана модель в форме (6.3). Будем считать, что модель адекватна, то есть достаточно точно аппроксимирует генеральное уравнение регрессии (ту идеальную математическую модель, которая точно описывает изучаемое явление). Для оценивания параметров модели используем независимую выборку объема n, каждый элемент которой (строка матрицы данных) содержит m значений факторов $x_1,\ x_2,\ \ldots,\ x_m$ и одно значение отклика y (всего $m+1$ значение в строке матрицы данных). В соответствии с моделью для каждого i-го элемента выборки можно записать так называемое **условное уравнение**

$$y_i = b_0 + b_1\,x_{i1} + \ldots + b_m\,x_{im} + e_i\,, \quad i = 1,\ \ldots,\ n, \qquad (6.5)$$

где $b_0,\ b_1,\ \ldots,\ b_m$ — *оценки* искомых неизвестных параметров $\beta_0,\ \beta_1,\ \ldots,\ \beta_m$; e_i — остаточное отклонение, или **невязка** условного уравнения.

Определение 6.4. ***Система условных уравнений*** *в матричной записи имеет вид*

$$\vec{\mathbf{Y}} = \hat{\mathrm{X}}\,\vec{\mathbf{B}} + \vec{\mathbf{E}}\,, \tag{6.6}$$

где $\vec{\mathbf{Y}}$ *— n-мерный вектор значений отклика;* $\hat{\mathrm{X}}$ *— матрица размера* $n \times (m+1)$*, у которой первый столбец состоит из единиц и соответствует свободному члену модели, а остальные представляют наблюдаемые значения факторов;* $\vec{\mathbf{B}}^T = (b_0,\ b_1,\ \ldots,\ b_m)$ *— искомый неизвестный вектор оценок параметров регрессионной модели;* $\vec{\mathbf{E}}^T = (e_1,\ e_2,\ \ldots,\ e_n)$ *— вектор остаточных отклонений, имеющий нулевое математическое ожидание и диагональную ковариационную матрицу* $\hat{\Sigma}_{\vec{\mathbf{E}}}$*, равную* $\sigma_0^2\,\hat{\mathrm{I}}$*, так как мы рассматриваем независимую выборку, а условные уравнения полагаем равноточными.*

Система уравнений (6.6) избыточна, то есть содержит больше уравнений, чем неизвестных, поскольку для получения в присутствии невязок точных оценок параметров $\beta_0,\ \ldots,\ \beta_m$ всегда следует придерживаться требования $n \gg m+1$. Для решения избыточной системы используют определенное условие, в качестве которого в регрессионном анализе за редким исключением используют принцип наименьших квадратов (далее для краткости — МНК). В соответствии с принципом МНК за наилучшее принимают решение, минимизирующее сумму квадратов остаточных отклонений:

$$SS_0 = \left(\vec{\mathbf{Y}} - \hat{\mathrm{X}}\vec{\mathbf{B}}\right)^T \left(\vec{\mathbf{Y}} - \hat{\mathrm{X}}\vec{\mathbf{B}}\right) = \vec{\mathbf{E}}^T\vec{\mathbf{E}} = \sum_{i=1}^{n} e_i^2\,. \tag{6.7}$$

Чтобы найти минимум функции (6.7), дифференцируем SS_0 по вектору $\vec{\mathbf{B}}$ и, приравнивая производную нулю, получаем **систему** так называемых **нормальных уравнений**

$$\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)\vec{\mathbf{B}} = \hat{\mathrm{X}}^T\vec{\mathbf{Y}}\,. \tag{6.8}$$

Если квадратная симметричная матрица $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$ не вырождена (ее определитель отличен от нуля), решение системы (6.8) можно получить, умножив левую и правую части (6.8) на матрицу $\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1}$, обратную к матрице $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$. Получаем решение

$$\vec{\mathbf{B}} = \left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1}\hat{\mathrm{X}}^T\vec{\mathbf{Y}}\,, \tag{6.9}$$

$$\check{\vec{\mathbf{Y}}} = \hat{\mathrm{X}}\,\vec{\mathbf{B}}\,, \tag{6.10}$$

где векторное уравнение (6.10) — набор уравнений линейной регрессии для всех строк матрицы данных.

Определение 6.5. *Матрица вида* $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$ *широко используется в многомерной статистике и называется* ***информационной матрицей****. Она симметрична, а элементы главной диагонали у нее положительны, так как каждый из них представляет собой сумму квадратов значений соответствующего фактора.*

МНК-решение (6.9) обладает рядом ценных свойств:

— *линейностью* по отношению к вектору $\vec{\mathbf{Y}}$;

— *несмещенностью*;

— *эффективностью*;

— *независимостью от закона распределения вектора откликов* $\vec{\mathbf{Y}}$.

Линейность по отношению к вектору $\vec{\mathbf{Y}}$ полезна в том смысле, что если для данного набора факторов у нас существует несколько реализаций вектора $\vec{\mathbf{Y}}$, то решить задачу отыскания параметров очень легко, так как не надо каждый раз пересчитывать коэффициент перед $\vec{\mathbf{Y}}$ в выражении (6.9), то есть для всех $\vec{\mathbf{Y}}$ самая трудоемкая часть решения проводится один раз.

При общей справедливости четвертого свойства наибольшая эффективность оценки параметров линейной регрессии достигается при нормальном распределении ошибок.

В прил. 2 приведены способы построения линейной регрессии методом наименьших квадратов.

6.3. Дисперсионный анализ уравнения регрессии

На рис. 6.1 в качестве примера показана выборочная (оцененная по принципу МНК) линия регрессии для случая простейшей линейной однофакторной модели и выделена одна из точек выборки $(x_i,\ y_i)$. Символом $\check{y}_i$ обозначено вычисленное по полученному уравнению регрессии ожидаемое (предсказанное) значение отклика при $x = x_i$. Здесь, как в случае дисперсионного анализа (см. предыдущую главу), можно выделить три отклонения: **общее отклонение** $y_i - \langle y \rangle$, где $\langle y \rangle$ — среднее арифметическое всех значений отклика в выборке; **остаточное отклонение** $y_i - \check{y}_i$ и отклонение, **обусловленное регрессией** $\check{y}_i - \langle y \rangle$, причем, как видно из рисунка, общее отклонение есть сумма остаточного отклонения и отклонения, обусловленного регрессией. Для сумм квадратов этих отклонений можно

записать

$$\underbrace{\sum_{i=1}^{n}(y_i - \langle y\rangle)^2}_{SS_{\vec{\mathbf{Y}}}\text{ — общее}} = \underbrace{\sum_{i=1}^{n}(\check{y}_i - \langle y\rangle)^2}_{SS_R\text{ — регрессионное}} + \underbrace{\sum_{i=1}^{n}(y_i - \check{y}_i)^2}_{SS_0\text{ — остаточное}} . \qquad (6.11)$$

Обозначим эти суммы соответственно $SS_{\vec{\mathbf{Y}}}$, SS_R и SS_0. В матричных обозначениях

$$SS_{\vec{\mathbf{Y}}} = \vec{\mathbf{Y}}^T\vec{\mathbf{Y}} - n\langle y\rangle^2 , \qquad (6.12)$$

$$SS_R = \check{\vec{\mathbf{Y}}}^T\check{\vec{\mathbf{Y}}} - n\langle y\rangle^2 , \qquad (6.13)$$

и в соответствии с дисперсионным соотношением (6.11)

$$\begin{aligned} SS_0 = SS_{\vec{\mathbf{Y}}} - SS_R = \vec{\mathbf{Y}}^T\vec{\mathbf{Y}} - \check{\vec{\mathbf{Y}}}^T\check{\vec{\mathbf{Y}}} = \\ = \vec{\mathbf{Y}}^T\vec{\mathbf{Y}} - \vec{\mathbf{B}}^T\hat{\mathrm{X}}^T\hat{\mathrm{X}}\,\vec{\mathbf{B}} = \vec{\mathbf{Y}}^T\vec{\mathbf{Y}} - \vec{\mathbf{B}}^T\hat{\mathrm{X}}^T\vec{\mathbf{Y}}, \end{aligned} \qquad (6.14)$$

где использованы следующие из матричной записи нормальных уравнений и их решения соотношения $\check{\vec{\mathbf{Y}}} = \hat{\mathrm{X}}\,\vec{\mathbf{B}}$ и $\hat{\mathrm{X}}^T\hat{\mathrm{X}}\,\vec{\mathbf{B}} = \hat{\mathrm{X}}^T\vec{\mathbf{Y}}$. Соотношение (6.14) удобно использовать для вычисления SS_0, так как требуется меньше затрат, чем прямое суммирование квадратов остаточных отклонений и меньше подвержено влиянию ошибок округления. Из (6.14) также видно, что $\check{\vec{\mathbf{Y}}}^T\check{\vec{\mathbf{Y}}} = \vec{\mathbf{B}}^T\hat{\mathrm{X}}^T\vec{\mathbf{Y}}$, то есть

$$SS_R = \vec{\mathbf{B}}^T\hat{\mathrm{X}}^T\vec{\mathbf{Y}} - n\langle y\rangle^2 . \qquad (6.15)$$

Результаты проведенных выкладок обычно сводят в **таблицу дисперсионного анализа** (табл. 6.1), где в последнем столбце приведены средние квадраты отклонений, но *толь-*

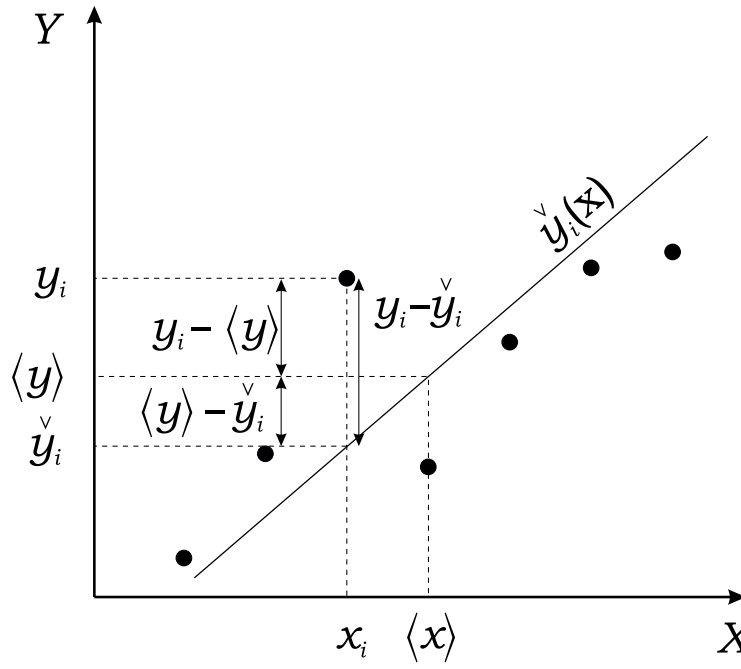

Рис. 6.1. Линейная регрессия и отклонения точек выборки от линии регрессии и от средних по выборке значений

ко один из них (остаточный) является дисперсией, поскольку два остальных связаны с неслучайной зависимостью отклика от факторов. Тем не менее все **средние квадраты** $MS_{\vec{\mathbf{Y}}}$, MS_R, s_0^2 (суммы квадратов, деленные на число степеней свободы) можно рассматривать как характеристики рассеяния соответственно значений y_i относительно $\langle y \rangle$, предсказанных по найденной регрессионной зависимости значений $\check{y}_i$ относительно $\langle y \rangle$ и значений $\check{y}_i$ относительно линии регрессии.

Таблица 6.1

Дисперсионный анализ уравнения регрессии

Сумма квадратов	Степени свободы	Средние
$SS_R = \vec{\mathbf{B}}^T \hat{\mathrm{X}}^T \vec{\mathbf{Y}} - n \langle y \rangle^2$	m	$MS_R = \dfrac{SS_R}{m}$
$SS_0 = \vec{\mathbf{Y}}^T \vec{\mathbf{Y}} - \vec{\mathbf{B}}^T \hat{\mathrm{X}}^T \vec{\mathbf{Y}}$	$n - m - 1$	$s_0^2 = \dfrac{SS_0}{n - m - 1}$
$SS_{\vec{\mathbf{Y}}} = \vec{\mathbf{Y}}^T \vec{\mathbf{Y}} - n \langle y \rangle^2$	$n - 1$	$MS_{\vec{\mathbf{Y}}} = \dfrac{SS_{\vec{\mathbf{Y}}}}{n - 1}$

На основании данных табл. 6.1 можно найти оценку множественного коэффициента корреляции R (R^2 — оценка выборочного коэффициента детерминации):

$$R^2 = \frac{SS_R}{SS_{\vec{\mathbf{Y}}}} = \frac{\vec{\mathbf{B}}^T \hat{\mathrm{X}}^T \vec{\mathbf{Y}} - n \langle y \rangle^2}{\vec{\mathbf{Y}}^T \vec{\mathbf{Y}} - n \langle y \rangle^2} \,. \tag{6.16}$$

Из рис. 6.1 и дисперсионного соотношения (6.11) видно, что *если бы все точки* $(x_i,\ y_i)$ *лежали на линии регрессии*, то SS_R совпала бы с $SS_{\vec{\mathbf{Y}}}$, а R^2 было бы равным единице. С другой стороны, *при независимости* математического ожидания отклика от факторов все коэффициенты β_k, кроме β_0, равны нулю, линия регрессии параллельна оси абсцисс, то есть $SS_{\vec{\mathbf{Y}}}$ должна *незначимо отличаться* от SS_0, а SS_R и R^2 — от нуля.

Определение 6.6. *Множественный коэффициент корреляции* R^2 *в виде* (6.16) *служит мерой близости связи между* $\vec{\mathbf{X}}$ *и* $\vec{\mathbf{Y}}$ *к функциональной (не обязательно линейной) зависимо-*

сти, а его численное значение, часто выражаемое в процентах, показывает, какую долю общего рассеяния значений y_i относительно $\langle y \rangle$ можно объяснить регрессией.

Значимость регрессии, то есть гипотезу

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0\,, \tag{6.17}$$

в случае нормального распределения случайной величины $\vec{\mathbf{Y}}$ проверяют, как и в дисперсионном анализе, по F-критерию со статистикой

$$F = \frac{MS_R}{s_0^2} = \frac{\vec{\mathbf{B}}^T \hat{\mathrm{X}}^T \vec{\mathbf{Y}} - n\,\langle y \rangle^2}{m\,s_0^2}\,, \tag{6.18}$$

критической областью для которого будет $F > F_\alpha(m,\ n-m-1)$, где $F_\alpha(m,\ n-m-1)$ — α-процентная точка распределения Фишера; $\alpha = 1 - \gamma$; γ — выбранная доверительная вероятность. При выполнении последнего неравенства регрессия значима, то есть по крайней мере один из факторов значимо влияет на математическое ожидание отклика. Какие именно из коэффициентов регрессии оказались значимыми, можно найти с помощью так называемого **частного F-критерия** (см. (6.19)).

Пусть нам надо проверить на значимость не все, а только последние l факторов, при этом, естественно, путем перестановки столбцов матрицы данных любые факторы можно сделать последними. Для проверки сначала найдем МНК-решения для двух моделей: полной m-факторной и содержащей $(m-l)$ факторов, когда последние l факторов не включены в модель. Пусть значения обусловленных регрессией сумм квадратов для двух решений оказались равными $SS_{R,\,m}$ и $SS_{R,\,m-l}$

соответственно. Разность этих сумм (обозначим ее $SS_{l|m}$), имеющую $m-(m-l)=l$ степеней свободы, называют **дополнительной суммой квадратов**. Если последние l факторов незначимы, то, как показывается в математической статистике, математическое ожидание отношения $SS_{l|m}/\,l$ равно σ_0^2, поэтому в качестве статистики критерия можно использовать отношение

$$F_{l|m}=\frac{SS_{l|m}}{l\,s_0^2}\,, \tag{6.19}$$

которое и называют **частным F-критерием**. Величина s_0^2 оценивается по полной m-факторной модели.

Критическая область критерия задается неравенством $F_{l|m}>F_\alpha(l,\ n-m-l)$, при выполнении которого коэффициенты $\beta_{m-l-1},\ \ldots,\ \beta_m$ следует считать значимыми. Чаще всего частный F-критерий применяется раздельно к каждому фактору, то есть для случая $l=1$, что помогает включить в модель все значимые факторы и отсеять незначимые (см. следующую главу).

Проверка значимости факторов необходима всегда, когда мы не уверены в выборе математической модели регрессии, а также в тех случаях, когда некоторые полученные нами оценки коэффициентов модели оказываются близкими по величине к их ошибкам.

6.4. Оценка точности решения МНК

При обсуждении вопроса об оценке точности решения МНК мы должны иметь в виду две стороны задачи: точность оценки полученного вектора параметров $\vec{\mathbf{B}}$, которая задается его

ковариационной матрицей, и точность предсказания значения отклика, характеризуемая дисперсией величины $\check{y}$.

Найдем сначала оценку ковариационной матрицы вектора $\vec{\mathbf{B}}$. Как видно из решения нормальных уравнений (6.9), вектор $\vec{\mathbf{B}}$ является линейной функцией вектора выборочных значений отклика $\vec{\mathbf{Y}}$, поэтому можно применить формулу линейного преобразования ковариационной матрицы (1.31), учтя при этом, что $\hat{\Sigma}_{\vec{\mathbf{Y}}} = \sigma_0^2\,\hat{\mathrm{I}}$:

$$\begin{aligned}\hat{\Sigma}_{\vec{\mathbf{B}}} &= \left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1}\hat{\mathrm{X}}^T\hat{\Sigma}_{\vec{\mathbf{Y}}}\hat{\mathrm{X}}\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1} = \\ &= \left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1}\hat{\mathrm{X}}^T\sigma_0^2\,\hat{\mathrm{I}}\,\hat{\mathrm{X}}\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1} = \sigma_0^2\,\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1},\end{aligned} \tag{6.20}$$

или заменяя дисперсию ее выборочной оценкой:

$$\check{\hat{\Sigma}}_{\vec{\mathbf{B}}} = \hat{\mathrm{S}}_{\vec{\mathbf{B}}} = s_0^2\,\left(\hat{\mathrm{X}}^T\hat{\mathrm{X}}\right)^{-1}. \tag{6.21}$$

Таким образом, ковариационная матрица вектора оценок коэффициентов регрессионной модели есть матрица, обратная к матрице коэффициентов системы нормальных уравнений и нормированная на s_0^2. *При нормальном распределении* $\vec{\mathbf{Y}}$ можно построить доверительную область для вектора $\vec{\boldsymbol{\beta}}$:

$$\left(\vec{\mathbf{B}} - \vec{\boldsymbol{\beta}}\right)^T \hat{\mathrm{S}}_{\vec{\mathbf{B}}}^{-1}\left(\vec{\mathbf{B}} - \vec{\boldsymbol{\beta}}\right) \leqslant F_\alpha(m+1,\ n-m-1)\,, \tag{6.22}$$

где $F_\alpha(m+1,\ n-m-1)$ — α-процентная точка распределения Фишера; $\alpha = 1-\gamma$; γ — выбранная доверительная вероятность.

Можно построить, что делают чаще, и одномерные доверительные интервалы раздельно для каждого β_i, записав их

в знакомом и привычном виде:

$$b_i \pm s_i\, t_\gamma(n-m-1)\,, \tag{6.23}$$

где s_i — среднеквадратическое отклонение для оценки b_i, равное квадратному корню из jj-го элемента ковариационной матрицы $\hat{\mathrm{S}}_{\vec{\mathbf{B}}}$; $t_\gamma(n-m-1)$ — γ-процентная точка распределения Стьюдента. Одномерные интервалы (6.23) используются для указания диапазонов возможных значений одного параметра безотносительно к значениям других. Неверно пытаться интерпретировать эти интервалы совместно, считая задаваемый соотношениями (6.23) $(m+1)$-мерный параллелепипед как совместную доверительную область. На рис. 6.2 приведены совместная эллиптическая доверительная область для вектора $\langle\vec{\boldsymbol{\beta}}\rangle = (\beta_0,\ \beta_1)$, а также раздельные доверительные интервалы $\Delta\beta_0$ и $\Delta\beta_1$. Из рисунка видно, что *попадание вектора $\vec{\boldsymbol{\beta}}$ в прямоугольник еще не гарантирует его попадания в истинную совместную доверительную область.*

Если совместная доверительная область (6.22) при разумно выбранных значениях доверительной вероятности включает начало координат, то истинные значения параметров $(\beta_0,\ \ldots,\ \beta_m)$ могут быть все равны нулю, так что регрессия, возможно, незначима. В таком случае необходимо провести статистическую проверку гипотезы (6.17) о значимости регрессии.

Если уравнение регрессии предназначено для использования в качестве эмпирической формулы для вычисления значений отклика, то ковариационная матрица нужна только для оценки точности предсказанных значений отклика $\check{y}$.

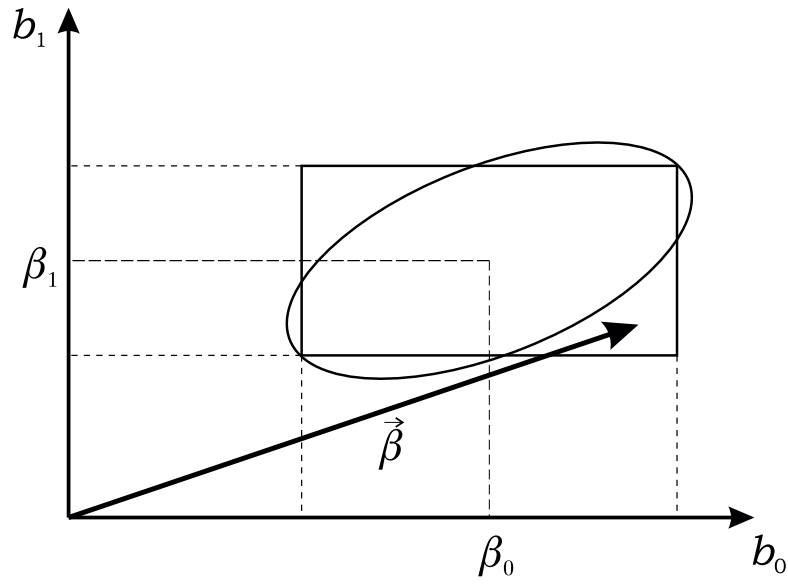

Рис. 6.2. Совместная доверительная область и раздельные доверительные интервалы для двумерного случайного вектора

При вычислении значений отклика для заданных значений факторов $\vec{\mathbf{X}}_*^T = (x_1, \ \ldots, \ x_m)$ по полученному уравнению регрессии величины $\check{y}$ являются линейной комбинацией случайных величин $\vec{\mathbf{B}}^T = (b_0, \ \ldots, \ b_m)$. Отсюда следует, что распределение оценки $\check{y}$ совпадает с распределением остаточных отклонений с точностью до постоянных параметров (см. главу 1 о линейных преобразованиях случайного вектора), а также то, что $\check{y}$ есть несмещенная оценка, так как $\hat{\mathrm{E}}\,(\check{y}) = \hat{\mathrm{E}}\left(\vec{\mathbf{X}}_*^T\vec{\mathbf{B}}\right) = \vec{\mathbf{X}}_*^T\hat{\mathrm{E}}\left(\vec{\mathbf{B}}\right) = \vec{\mathbf{X}}_*^T\vec{\boldsymbol{\beta}}$. Кроме того,

дисперсия оценки $\check{y}$ равна, что следует из (1.31) и (6.21),

$$s^2_{\check{y}} = \vec{\mathbf{X}}^T_* \hat{\mathrm{S}}_{\vec{\mathbf{B}}} \vec{\mathbf{X}}_* = s^2_0 \, \vec{\mathbf{X}}^T_* \left(\hat{\mathrm{X}}^T \hat{\mathrm{X}} \right)^{-1} \vec{\mathbf{X}}_* \, . \tag{6.24}$$

Следует отметить, что величина $\check{y}$ является предсказанным значением математического ожидания отклика $\hat{\mathrm{E}} \left(y \left(\vec{\mathbf{X}}_* \right) \right)$, а не самой случайной величины Y, которая имеет дисперсию σ^2_0. При неограниченном увеличении объема выборки и при адекватной модели точность предсказания математического ожидания будет тоже неограниченно возрастать, однако очевидно, что предсказать индивидуальное значение отклика точнее, чем с дисперсией σ^2_0, невозможно. Таким образом, к величине, определяемой формулой (6.24), следует прибавить величину s^2_0, так что окончательно для полной дисперсии отклика при учете всех влияний ошибок имеем

$$s^2_{\mathrm{total},\, \check{y}} = s^2_0 \left(1 + \vec{\mathbf{X}}^T_* \left(\hat{\mathrm{X}}^T \hat{\mathrm{X}} \right)^{-1} \vec{\mathbf{X}}_* \right) . \tag{6.25}$$

Последнее можно понимать так, что полная ошибка предсказанного значения определяется как ошибкой, вызываемой дисперсией откликов, так и ошибками оценок коэффициентов регрессионной модели. Доверительный интервал для предсказанного значения отклика можно найти по формуле (6.23), в которой величины b_i и s_i следует заменить на $\check{y}$ и $s_{\check{y}}$ соответственно, либо на $\check{y}$ и $s_{\mathrm{total},\, \check{y}}$ при учете дисперсии откликов.

6.5. Анализ главных компонент

Широкое распространение в статистических исследованиях получили так называемые **факторные методы**, одним из которых является уже рассматривавшийся в главе 5 дисперсионный анализ. Дисперсионный анализ, как мы видели, позволяет выделить факторы, значимо влияющие на математическое ожидание случайной величины. Однако дисперсионный анализ рассчитан в основном на *активный эксперимент*. Но для звездной астрономии характерен пассивный эксперимент, так как наблюдатель не может выбирать способ воздействия на изучаемый объект. Для пассивного эксперимента разработан ряд специальных методов, таких как **дискриминантный анализ**, **кластерный анализ**, **факторный анализ** с его разновидностью — **анализом главных компонент**. Встречается и другое название последнего метода — **регрессия на главных компонентах**, что подчеркивает связь этого метода с регрессионным анализом. Данный метод уже упоминался в главе 1. Остановимся подробнее на этом, наиболее известном из факторных методов.

Пусть имеется случайный вектор $\vec{\mathbf{X}}^T = (x_1, \ldots, x_m)$ с математическим ожиданием $\vec{\boldsymbol{\xi}}$ и ковариационной матрицей $\hat{\Sigma}_{\vec{\mathbf{X}}}$, которой соответствует гиперэллипсоид рассеяния (1.17). Во многих практических случаях можно предположить, что существует некоторое число $k < m$ *непосредственно не наблюдаемых некоррелированных случайных переменных*, совместное распределение которых имеет приблизительно такой же по форме и размерам гиперэллипсоид рассеяния в подпространстве размерности k, что и вектор $\vec{\mathbf{X}}$. Это возможно при сильной корреляции некоторых из компонент вектора $\vec{\mathbf{X}}$, вследствие чего

матрица $\hat{\Sigma}_{\vec{\mathbf{X}}}$ близка к вырожденной. В этом случае мы можем свести m-мерное распределение к распределению меньшей размерности k и выдвинуть новые гипотезы, с иных позиций интерпретировать поведение вектора $\vec{\mathbf{X}}$. В других случаях этот метод перехода к новому случайному вектору меньшей размерности с некоррелированными координатами создает бóльшие вычислительные удобства, в частности, для проведения регрессионного анализа. Переход к новым переменным можно осуществить на основе анализа ковариационной матрицы.

Определение 6.7. *В методе главных компонент новыми переменными являются линейные комбинации компонент случайного вектора* $\vec{\mathbf{X}}$*, вычисляемые по формуле* (1.37)

$$\vec{\mathbf{Z}} = \hat{\mathrm{V}}^T \left(\vec{\mathbf{X}} - \vec{\boldsymbol{\xi}} \right) , \tag{6.26}$$

где $\vec{\mathbf{Z}}$ *— преобразованный случайный вектор;* $\hat{\mathrm{V}}$ *—* $(m \times m)$*-матрица преобразования, столбцы которой являются собственными векторами матрицы* $\hat{\Sigma}_{\vec{\mathbf{X}}}$*.*

Главными компонентами называются координаты z_k *случайного вектора* $\vec{\mathbf{Z}}$*, получающегося после преобразования вектора* $\vec{\mathbf{X}}$*, и* $\vec{\mathbf{V}}_k$*, — собственные векторы матрицы* $\hat{\Sigma}_{\vec{\mathbf{X}}}$*.*

Напомним, что ковариационная матрица $\hat{\Sigma}_{\vec{\mathbf{Z}}}$ вектора $\vec{\mathbf{Z}}$ диагональна с элементами главной диагонали — дисперсиями главных компонент, равными собственным значениям $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_m$ матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$. Напомним также, что произведение собственных значений равно определителю матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$,

а их сумма равна следу, то есть сумме диагональных элементов матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$. Таким образом, след ковариационной матрицы и ее определитель являются инвариантами преобразования (6.26).

Если распределение вырождено и случайный вектор $\vec{\mathbf{X}}$ является в действительности не m-мерным, а k-мерным, то $m-k$ собственных значений ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$ равны нулю, а ее ранг равен k.

Рассмотрим теперь m-мерную выборку объема n, заданную в форме матрицы данных с m столбцами и n строками. Найдем сначала оценку ковариационной матрицы $\hat{\Sigma}_{\vec{\mathbf{X}}}$, то есть матрицу $\hat{\mathrm{S}}_{\vec{\mathbf{X}}}$, а затем одним из аналитических или численных методов найдем ее собственные значения $l_1,\ \ldots,\ l_m$ и собственные векторы $\vec{\mathbf{V}}_1,\ \ldots,\ \vec{\mathbf{V}}_m$, после чего выборочные значения главных компонент определятся формулой

$$\hat{\mathrm{Z}} = \hat{\mathrm{X}}^0\,\hat{\mathrm{V}}\,, \tag{6.27}$$

где $\hat{\mathrm{X}}^0$ — матрица центрированных исходных данных, элементами которой являются величины $x^0_{ik} = x_{ik} - \langle x\rangle_k$; $\hat{\mathrm{V}}$ — матрица, столбцами которой являются собственные векторы выборочной ковариационной матрицы $\hat{\mathrm{S}}_{\vec{\mathbf{X}}}$; $\hat{\mathrm{Z}}$ — $(n\times m)$-матрица преобразованных данных или матрица выборочных значений главных компонент.

И для нашей выборки $\hat{\mathrm{S}}_{\vec{\mathbf{Z}}}$ диагональна с элементами главной диагонали $l_1,\ \ldots,\ l_m$, являющимися выборочными дисперсиями главных компонент $z_1,\ z_2,\ \ldots,\ z_m$. Очевидно, что даже если некоторые из генеральных дисперсий главных компонент равны нулю, их выборочные значения не окажутся нулевыми,

поэтому в случае выборки речь может идти не о равенстве нулю некоторых собственных значений, а о их *незначимости*, то есть задача уменьшения размерности вектора $\vec{\mathbf{X}}$ сводится к проверке гипотез вида

$$H_0 : \lambda_1 = \ldots = \lambda_m \quad \text{или} \quad H_0 : \lambda_j = 0\,. \tag{6.28}$$

Первая из них утверждает, что гиперэллипсоид рассеяния случайного вектора $\vec{\mathbf{Z}}$, а значит и вектора $\vec{\mathbf{X}}$, является гиперсферой, то есть что координаты случайного вектора $\vec{\mathbf{X}}$ независимы и имеют равные дисперсии. Ясно, что если эта гипотеза не отвергается, то размерность распределения не может быть уменьшена. Статистикой для проверки первой гипотезы может служить величина

$$\chi^2 = -(n-1)\left(\sum_{j=1}^{m} \ln l_j - m\,\ln\left(\frac{1}{m}\sum_{j=1}^{m} l_j\right)\right), \tag{6.29}$$

которая при нормальном распределении случайного вектора $\vec{\mathbf{X}}$ и при большом объеме выборки распределена в соответствии с законом Пирсона с $(m(m+1)/2) - 1$ степенями свободы. Гипотеза отвергается, если величина χ^2 превысит критическое значение χ^2_α с указанным числом степеней свободы.

Вторую гипотезу обычно формулируют не для одной, а сразу для нескольких величин λ_j, то есть в форме $H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots \lambda_m = 0$, при этом предполагается, что выборочные собственные значения матрицы $\hat{\mathrm{S}}_{\vec{\mathbf{X}}}$ упорядочены, то есть $l_1 \geqslant l_2 \geqslant \ldots \geqslant l_m$. Для приближенной оценки величины k — номера последнего собственного значения, предпола-

гаемого значимым, — следует вычислить долю вклада первых k собственных значений в суммарную дисперсию, то есть найти величину отношения $(l_1 + \ldots + l_k)/(l_1 + \ldots + l_m)$ для ряда значений $k = 1,\ 2$ и т. д. и остановиться на том значении k, при котором эта доля станет достаточно близкой к единице, например, большей, чем 0.8 или 0.9. Проверку значимости собственных значений по статистическим критериям обычно не проводят, так как эти критерии в данном случае приносят мало пользы и весьма чувствительны к предположению о нормальности распределения случайного вектора $\vec{\mathbf{X}}$. Вполне достаточно построить график величин собственных значений в зависимости от их номера.

Возможность уменьшения размерности случайного вектора и некоррелированность главных компонент делает рассматриваемый метод весьма привлекательным для использования в регрессионном анализе. Пусть задана система условных уравнений $\vec{\mathbf{Y}} = \hat{\mathrm{X}}\,\vec{\mathbf{B}}$, которую можно переписать в центрированной форме $\vec{\mathbf{Y}}^0 = \hat{\mathrm{X}}^0\,\vec{\mathbf{B}}$, где элементами являются центрированные величины $y_i^0 = y_i - \langle y \rangle$ и $x_{ij}^0 = x_{ij} - \langle x \rangle_j$ соответственно. Как обычно, в случае центрированных переменных регрессионная модель не будет содержать свободного члена β_0 и размер матрицы $\hat{\mathrm{X}}^0$ будет $n \times m$. Найдем выборочную ковариационную матрицу $\hat{\mathrm{S}}_{\vec{\mathbf{B}}}$, вычислим для нее собственные значения $l_1,\ \ldots,\ l_m$, собственные векторы $\vec{\mathbf{V}}_1,\ \ldots,\ \vec{\mathbf{V}}_m$ и преобразуем матрицу $\hat{\mathrm{X}}^0$ в матрицу выборочных значений главных компонент $\hat{\mathrm{Z}}$ по формуле (6.27). Если среди собственных значений окажутся незначимые, то из матрицы $\hat{\mathrm{Z}}$ можно исключить соответствующие столбцы, после чего ее размер уменьшится до $n \times k$, где k — число оставшихся главных компонент. Условные и нормальные

уравнения теперь соответственно примут вид

$$\begin{aligned} \vec{\mathbf{Y}}^0 &= \hat{\mathrm{Z}}\vec{\mathbf{A}}\,, \\ \hat{\mathrm{Z}}^T\hat{\mathrm{Z}}\vec{\mathbf{A}} &= \hat{\mathrm{Z}}^T\vec{\mathbf{Y}}\,, \end{aligned} \tag{6.30}$$

где $\vec{\mathbf{A}}$ — $(k \times 1)$-вектор новых коэффициентов регрессии. Поскольку главные компоненты некоррелированы, матрица $\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}$ оказывается диагональной с элементами главной диагонали $(n-1)\,l_j$. Обратная матрица $\left(\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}\right)^{-1}$ тоже будет диагональной с элементами $((n-1)\,l_j)^{-1}$, поэтому оценки коэффициентов регрессии и их дисперсии можно найти непосредственно по формулам

$$\begin{aligned} \alpha_j &= \frac{1}{(n-1)\,l_j}\sum_{i=1}^{n} z_{ij}\,y_i\,, \\ s^2_{\alpha_j} &= \frac{s_0^2}{(n-1)\,l_j}\,. \end{aligned} \tag{6.31}$$

Так как оценки коэффициентов α_j некоррелированы, то их значимость можно проверить по t-критерию (6.23), не прибегая к построению многомерной критической области.

Теперь следует сделать одно весьма существенное замечание относительно метода главных компонент. Дело в том, что главные компоненты *не инвариантны относительно изменения масштаба или единиц измерения компонент случайного вектора* $\vec{\mathbf{X}}$. Изменение масштаба хотя бы одной компоненты вектора $\vec{\mathbf{X}}$, скажем в десять раз, приведет к изменению дисперсии этой компоненты в сто раз, что неизбежно скажется на дисперсиях главных компонент и на их значимости. Поэтому метод главных компонент, строго говоря, применим толь-

ко к тем случайным векторам, компоненты которых являются однородными случайными величинами, измеренными в одних и тех же единицах. Поэтому изложенный метод эффективно используется для задач калибровки, когда характеристики звезд, такие как абсолютные звездные величины, избытки цвета или металличности $[\,\mathrm{Fe/H}\,]$, связываются с показателями цвета какой-либо фотометрической системы.

Иногда неоднородность величин пытаются преодолеть переходом к стандартизованным переменным $u_j = \left(x_j - \left\langle \vec{\mathbf{X}} \right\rangle_j\right)/\sigma_j$, но такой прием приводит к изменению соотношения полуосей гиперэллипсоида рассеяния, и применять его следует с осторожностью. Однако у данного преобразования есть и важное преимущество, так как в этом случае ковариационная матрица превращается в корреляционную со всеми преимуществами корреляционной матрицы.

7. ПРАКТИЧЕСКИЕ ВОПРОСЫ РЕГРЕССИОННОГО АНАЛИЗА

7.1. Взвешенный метод наименьших квадратов

В практическом применении регрессионного анализа иногда приходится иметь дело с выборкой, элементы которой имеют различную точность, то есть дисперсии условных уравнений неодинаковы. Может оказаться, что элементы выборки коррелированы, и тогда ковариационная матрица остаточных отклонений $\hat{\Sigma}_{\vec{\mathbf{E}}}$ уже не будет диагональной. Пусть в общем случае задана система условных уравнений (6.6)

$$\vec{\mathbf{Y}} = \hat{\mathrm{X}}\,\vec{\mathbf{B}} + \vec{\mathbf{E}}\,, \tag{7.1}$$

но $\hat{\Sigma}_{\vec{\mathbf{E}}} = \sigma_0^2\,\hat{\mathrm{V}} \neq \sigma_0^2\,\hat{\mathrm{I}}$, где $\hat{\mathrm{V}}$ — симметричная $(n \times n)$-матрица с положительными, не равными друг другу диагональными элементами.

Можно показать, что существует единственная невырожденная симметричная матрица $\hat{\mathrm{W}}$, такая, что $\hat{\mathrm{W}}^T\hat{\mathrm{W}} = \hat{\mathrm{W}}\,\hat{\mathrm{W}} = \hat{\mathrm{V}}$. Умножив систему уравнений (7.1)

слева на обратную матрицу $\hat{\mathrm{W}}^{-1}$ и введя обозначения

$$\vec{\mathbf{Y}}_0 \equiv \hat{\mathrm{W}}^{-1}\vec{\mathbf{Y}}\,, \tag{7.2}$$

$$\hat{\mathrm{X}}_0 \equiv \hat{\mathrm{W}}^{-1}\hat{\mathrm{X}}\,, \tag{7.3}$$

$$\vec{\mathbf{E}}_0 \equiv \hat{\mathrm{W}}^{-1}\vec{\mathbf{E}}\,, \tag{7.4}$$

получим новую систему уравнений

$$\vec{\mathbf{Y}}_0 = \hat{\mathrm{X}}_0\,\vec{\mathbf{B}} + \vec{\mathbf{E}}_0\,, \tag{7.5}$$

для которой матрица $\hat{\Sigma}_{\vec{\mathbf{E}}_0} = \sigma_0^2\,\hat{\mathrm{I}}$ уже диагональна. Действительно, так как преобразование $\vec{\mathbf{E}}_0 = \hat{\mathrm{W}}^{-1}\vec{\mathbf{E}}$ линейно, то, используя формулу (1.31), получаем $\hat{\Sigma}_{\vec{\mathbf{E}}_0} = \hat{\mathrm{W}}^{-1}\hat{\Sigma}_{\vec{\mathbf{E}}}\hat{\mathrm{W}}^{-1} = \sigma_0^2\,\hat{\mathrm{W}}^{-1}\,\hat{\mathrm{V}}\,\hat{\mathrm{W}}^{-1} = \sigma_0^2\,\hat{\mathrm{I}}$.

Таким образом, новые случайные величины $\vec{\mathbf{Y}}_0$ и $\vec{\mathbf{E}}_0$ оказываются равноточными и некоррелированными, а преобразованную систему условных уравнений (7.5) можно решать по обычным формулам регрессионного анализа.

Вид матрицы $\hat{\mathrm{W}}$ определить достаточно трудно, поэтому обычно ограничиваются случаем, когда эта матрица диагональна, то есть *когда условные уравнения неравноточны, но некоррелированы.* Матрица $\hat{\mathrm{W}}^{-1}$ будет в этом случае тоже диагональна, а ее ненулевыми элементами будут величины σ_0/σ_i, в чем можно убедиться, проделав обратный переход от $\hat{\Sigma}_{\vec{\mathbf{E}}_0} = \sigma_0^2\,\hat{\mathrm{I}}$ к матрице $\hat{\Sigma}_{\vec{\mathbf{E}}} = \sigma_0^2\,\hat{\mathrm{V}}$. Величину

$$w_i^2 = \frac{\sigma_0^2}{\sigma_i^2} \tag{7.6}$$

называют **весом** i-го условного уравнения.

Из сказанного следует известное правило: *чтобы сделать неравноточные условные уравнения равноточными, необходимо каждое условное уравнение исходной системы умножить на квадратный корень из его веса*.

На практике часто встречается ситуация, когда дисперсии условных уравнений σ_i^2 неизвестны, но имеются сведения, что значения отклика неравноточны. В этом случае за вес w_i^2 можно принять какую-либо величину, характеризующую точность данных наблюдений, например, число измерений данной величины, так как дисперсия результата при равноточной методике обратно пропорциональна числу измерений.

Полезно отметить, что если оценивать параметры $\langle\vec{\boldsymbol{\beta}}\rangle$ непосредственно по неравноточным условным уравнениям, то оценки b_j останутся несмещенными, но их дисперсии, а также дисперсии предсказанных значений окажутся большими, чем при применении взвешенного МНК.

7.2. Преобразование исходных данных: центрирование и нормирование

В регрессионных задачах матрица нормальных уравнений $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$ часто оказывается плохо обусловленной, то есть близкой к вырожденной. Часто такое положение встречается при приближении наблюдаемой зависимости полиномом. В этом случае ее определитель мал, и любое малое изменение, например округление, хотя бы одного из ее элементов приводит к значительному изменению обратной матрицы, а значит, и решения.

Одной из причин плохой обусловленности является резкое различие элементов матрицы $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$ по абсолютной величине, поэтому одним из способов борьбы с плохой обусловленностью может быть такое преобразование исходных данных, после которого матрица нормальных уравнений имела бы близкие по модулю элементы.

Рассмотрим одно из таких преобразований. Заменим в системе условных уравнений (6.6) величины y_i, x_{ik} и e_i на

$$u_i = \frac{y_i - \langle y \rangle}{\sqrt{SS_{\vec{\mathbf{Y}}}}}\,, \tag{7.7}$$

$$z_{ik} = \frac{x_{ik} - \langle x \rangle_k}{\sqrt{SS_k}}\,, \tag{7.8}$$

$$d_i = \frac{e_i}{\sqrt{SS_0}}\,, \tag{7.9}$$

где $\langle x \rangle_k$ — среднее арифметическое из элементов k-го столбца матрицы исходных данных $\hat{\mathrm{X}}$; SS_k — сумма квадратов отклонений этих элементов от их среднего. Легко заметить, что векторы $\vec{\mathbf{U}}^T = (u_1,\ u_2,\ \ldots,\ u_n)$ и $\vec{\mathbf{Z}}_k^T = (z_{1k},\ z_{2k},\ \ldots,\ z_{nk})$ имеют единичную длину, так как суммы квадратов их элементов равны единице. Системы условных и нормальных уравнений в новых переменных примут вид

$$\vec{\mathbf{U}} = \hat{\mathrm{Z}}\vec{\mathbf{A}} + \vec{\mathbf{D}}\,, \tag{7.10}$$

$$\left(\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}\right)\vec{\mathbf{A}} = \hat{\mathrm{Z}}^T\vec{\mathbf{U}}\,, \tag{7.11}$$

где $\vec{\mathbf{A}} = (a_1,\ a_2,\ \ldots,\ a_m)$ — вектор оценок новых, соответствующих преобразованным данным, коэффициентов регрес-

сии. Отметим, что в соответствии с (7.8) $z_{i\,0} = 0$, так как все $x_{i\,0}$ были равны единице, поэтому вектор $\vec{\mathbf{A}}$ не имеет элемента a_0, то есть в преобразованной системе координат поверхность регрессии проходит через ее начало (свободный член в регрессионной модели отсутствует). Соответственно размерности вектора $\hat{\mathrm{Z}}^T\vec{\mathbf{U}}$ и матриц $\hat{\mathrm{Z}}$ и $\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}$ уменьшены на единицу.

Можно показать, что элементами матрицы $\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}$ являются величины

$$r_{jk} = \frac{\sum\limits_{i=1}^{n} \left(x_{ij} - \langle x\rangle_j\right)\left(x_{ik} - \langle x\rangle_k\right)}{\sqrt{\sum\limits_{i=1}^{n} \left(x_{ij} - \langle x\rangle_j\right)^2 \sum\limits_{i=1}^{n} \left(x_{ik} - \langle x\rangle_k\right)^2}}\,, \tag{7.12}$$

то есть это выборочные парные коэффициенты корреляции j-го фактора с k-м, а элементами вектора $\hat{\mathrm{Z}}^T\vec{\mathbf{U}}$ являются выборочные коэффициенты корреляции отклика с каждым из факторов. Нормальные уравнения теперь можно записать, используя принятые здесь обозначения для корреляционной матрицы и вектора корреляций

$$\hat{\mathrm{R}}_{\vec{\mathbf{X}}}\vec{\mathbf{A}} = \vec{\mathbf{R}}_{\vec{\mathbf{Y}}}\,, \tag{7.13}$$

где $\vec{\mathbf{R}}_{\vec{\mathbf{Y}}} \equiv \hat{\mathrm{Z}}^T\vec{\mathbf{U}}$ называют **вектором корреляции**.

Поскольку коэффициенты корреляции по модулю не превосходят единицы и, с другой стороны, выборочные коэффициенты корреляции редко бывают близки к нулю, то элементы матрицы $\hat{\mathrm{Z}}^T\hat{\mathrm{Z}}$ оказываются величинами одного порядка, что сводит к минимуму ошибки округления.

После определения из нормальных уравнений компонент вектора $\vec{\mathbf{A}}$

$$\vec{\mathbf{A}} = \hat{\mathrm{R}}^{-1}_{\vec{\mathbf{X}}} \vec{\mathbf{R}}_{\vec{\mathbf{Y}}} \tag{7.14}$$

можно найти оценки коэффициентов исходной модели b_j по формулам

$$\begin{aligned} b_j &= a_j \sqrt{\frac{SS_{\vec{\mathbf{Y}}}}{SS_j}}\,, \\ b_0 &= y - \sum_{j=1}^{m} b_j\, x_j\,, \end{aligned} \tag{7.15}$$

а также и другие необходимые величины: суммы квадратов, квадрат множественного коэффициента корреляции R^2, функцию распределения F, выборочную ковариационную матрицу $S_{\vec{\mathbf{B}}}$.

7.3. Ошибки в факторах

В классическом МНК предполагается, что *все случайные погрешности наблюдений сосредоточены в величинах отклика*. В реальных задачах это условие часто *не выполняется*.

Рассмотрим задачу сведения двух каталогов собственных движений в один сводный. Для определенности эту задачу рассмотрим на примере рассеянного скопления, чтобы ограничиться небольшой областью неба. В этом случае мы можем выбрать один каталог и искать формулы приведения второго каталога к системе первого. Эти формулы в простейшем случае имеют

вид

$$\mu_{\alpha 1} = a_\alpha \, \mu_{\alpha 2} + b_\alpha \, \mu_{\alpha 2} + c_\alpha \, , \tag{7.16}$$

$$\mu_{\delta 1} = a_\delta \, \mu_{\delta 2} + b_\delta \, \mu_{\delta 2} + c_\delta \, , \tag{7.17}$$

где $\mu_{\alpha 1}$ и $\mu_{\delta 1}$ — соответственно компоненты собственного движения по прямому восхождению и склонению первого каталога, а индекс 2 обозначает собственные движения из второго каталога.

Элементы *матрицы преобразования* a_α, a_δ, b_α, b_δ и координаты *вектора смещения* c_α, c_δ необходимо оценить с помощью *общих* для двух каталогов звезд. Если бы отсутствовали ошибки в определении собственных движений, определитель матрицы преобразования (7.16), (7.17) был бы равен единице, так что задача сведения каталогов решалась бы путем введения постоянной поправки в собственные движения второго каталога и, возможно, к небольшому повороту системы координат. Однако, как было рассмотрено в главе 4, случайные ошибки размывают, растягивают плотность распределения, что ведет к увеличению в среднем величин собственных движений. В этом случае такое увеличение скажется в неравенстве определителя матрицы преобразования (7.16), (7.17) единице и его величина будет связана с величинами дисперсии случайных ошибок собственных движений первого и второго каталогов. Но этим не ограничивается влияние случайных ошибок на результаты оценивания параметров преобразования (7.16), (7.17), которое проводится с помощью регрессионного анализа.

Рассмотрим влияние ошибок в факторах на примере очень простой регрессионной модели:

$$y = a\,x\,. \tag{7.18}$$

Пусть ошибки ϵ_i присутствуют *только в величинах отклика*. Потребуем, чтобы ошибки не были коррелированы со значениями отклика и факторами, что обычно выполняется на практике. Регрессионная модель (условные уравнения) примет вид

$$y_i + \epsilon_i = a\,x_i\,, \tag{7.19}$$

где индекс i задает номер элемента выборки.

Получим выражение для оценивания коэффициента a, как делалось в предыдущей главе. Образуем невязки условных уравнений

$$\epsilon_i = y_i - a\,x_i \tag{7.20}$$

и сумму квадратов невязок (N — объем выборки)

$$SS = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - a\,x_i)^2\,. \tag{7.21}$$

Найдем значение параметра a, минимизирующее сумму квадратов SS:

$$\frac{\partial\,SS}{\partial a} = -2 \sum_{i=1}^{N} x_i\,(y_i - a\,x_i)\,. \tag{7.22}$$

Из этого выражения, раскрывая скобки, имеем

$$\sum_{i=1}^{N} x_i\, y_i - a \sum_{i=1}^{N} x_i^2 = 0\,. \tag{7.23}$$

Окончательно для оценки параметра регрессионной модели (7.18) получим выражение

$$\breve{a} = \frac{\sum\limits_{i=1}^{N} x_i\, y_i}{\sum\limits_{i=1}^{N} x_i^2}\,. \tag{7.24}$$

Как видим, в выражение (7.24) ошибки не входят, что доказывает независимость от распределения ошибок и его параметров процедуры оценивания методом наименьших квадратов коэффициента модели (7.18), но вид распределения ошибок может влиять на эффективность оценок.

Теперь рассмотрим случай, когда ошибки δ_i входят в значение фактора, причем *о распределении ошибок не делается никаких предположений*. В этом случае условные уравнения выглядят следующим образом:

$$y_i + \epsilon_i = a\,(x_i + \delta_i)\ . \tag{7.25}$$

Опуская промежуточные алгебраические преобразования, полностью аналогичные приведенным выше, получаем форму-

лу для оценивания параметра модели для данного случая:

$$\breve{a} = \frac{\sum\limits_{i=1}^{N} x_i\, y_i}{\sum\limits_{i=1}^{N} x_i^2 + \sum\limits_{i=1}^{N} \delta_i^2}\,. \tag{7.26}$$

В выражение (7.26) для оценивания параметра регрессионной модели непосредственно входит не равная нулю сумма квадратов ошибок факторов, и оценка параметра, если получать ее из выражения (7.24), оказывается *смещенной*.

Опишем кратко более общую модель с ошибками в факторах, которая для случая линейной m-факторной зависимости дается соотношением

$$\eta = \beta_0 + \sum_{i=1}^{m} \beta_i\, z_i\,. \tag{7.27}$$

Уравнение (7.27) связывает истинные значения факторов и отклика, которые, конечно, отягощены ошибками наблюдений. Вместо них наблюдения дают величины

$$\begin{aligned} y_k &= \eta_k + \gamma_k, \quad k = 1,\ 2,\ \ldots,\ N\,, \\ x_{ki} &= z_{ki} + \delta_{ki}, \quad i = 1,\ 2,\ \ldots,\ m\,, \end{aligned} \tag{7.28}$$

где η_k и z_{ki} — истинные значения отклика и факторов; γ_k и δ_{ki} — случайные ошибки отклика и факторов. Требуется по данным N измерений оценить неизвестные параметры $\beta_0,\ \ldots,\ \beta_m$.

Для наблюдаемых величин модель (7.27), (7.28) принимает вид

$$y_k = \beta_0 + \sum_{i=1}^{m} \beta_i \, x_{ki} + \epsilon_k \,, \tag{7.29}$$

где

$$\epsilon_k = \gamma_k - \sum_{i=1}^{m} \beta_i \, \delta_{ki} \tag{7.30}$$

представляет собой ошибки условных уравнений. Сделаем для простоты достаточно естественное предположение о равенстве нулю математического ожидания этой величины

$$\hat{\mathrm{E}}(\epsilon_k) = \hat{\mathrm{E}}(\gamma_k) - \sum_{i=1}^{m} \beta_i \, \hat{\mathrm{E}}(\delta_{ki}) = 0 \,, \tag{7.31}$$

а также одинаковости дисперсий

$$\begin{aligned} \sigma^2(\epsilon_k) = \hat{\mathrm{E}} \left(\left(\gamma_k - \sum_{i=1}^{m} \beta_i \, \delta_{ki} \right)^2 \right) = \\ = \sigma_\gamma^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} \beta_i \, \beta_j \, \rho_{ij} \, \sigma_{x_i} \, \sigma_{x_j} \,, \end{aligned} \tag{7.32}$$

где для краткости введены обозначения $\sigma_\gamma^2 \equiv \sigma^2(\gamma_k)$, $\sigma_{x_i}^2 \equiv \sigma^2(\delta_{ki})$ и $\rho_{ij} \, \sigma_{x_i} \sigma_{x_j} \equiv \mathrm{cov}\,(\delta_{ki}, \, \delta_{kj})$, причем ошибки модели предполагаются некоррелированными:

$$\begin{aligned} &\mathrm{cov}\,(\epsilon_k, \, \epsilon_j) = \\ &= \hat{\mathrm{E}} \left(\left(\gamma_k - \sum_{i=1}^{m} \beta_i \, \delta_{ki} \right) \left(\gamma_j - \sum_{i=1}^{m} \beta_i \, \delta_{ji} \right) \right) = 0 \,. \end{aligned} \tag{7.33}$$

В случае такой модели наблюдаемые значения факторов x_{ki} являются случайными величинами, коррелированными с ошибками модели

$$\begin{aligned}&\operatorname{cov}(x_{kj},\ \epsilon_k) = \\ &= \hat{\mathrm{E}}\left((z_{kj}+\delta_{kj})\left(\gamma_k+\beta_j\,\delta_{kj}+\sum_{i=1}^{m}\beta_i\,\delta_{ki}\right)\right) = -\beta_i\,\sigma^2_{x_i}\,.\end{aligned} \tag{7.34}$$

В этом и состоит основное различие между моделью с ошибками в факторах и стандартной регрессионной моделью. Последние выражения даны без выводов, которые можно найти в специальных статьях [22].

Вернемся к уже привычным матрично-векторным обозначениям. В этих обозначениях рассматриваемая нами модель (7.27)—(7.28) имеет вид

$$\vec{\mathbf{Y}} = \hat{\mathrm{F}}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\epsilon}}, \tag{7.35}$$

где вновь введены обозначения для погрешностей

$$\vec{\boldsymbol{\epsilon}} = \vec{\boldsymbol{\gamma}} - \hat{\Delta}\vec{\boldsymbol{\beta}}\,, \tag{7.36}$$

а матрица данных содержит ошибки

$$\hat{\mathrm{F}} = \hat{\mathrm{Z}} + \hat{\Delta}\,, \quad \hat{\Delta} = \left\{\begin{array}{cccc} 0 & \delta_{11} & \dots & \delta_{1m} \\ 0 & \delta_{21} & \dots & \delta_{2m} \\ \dots & \dots & \dots & \dots \\ 0 & \delta_{N1} & \dots & \delta_{Nm} \end{array}\right\}. \tag{7.37}$$

Как всегда, первый столбец соответствует свободному члену, так что в матрице данных первый столбец состоит из единиц, которые, естественно, имеют нулевые ошибки.

Из-за наличия указанной выше (см. формулу (7.34)) корреляции обычные МНК-оценки $\vec{\mathbf{B}} = \left(\hat{\mathrm{F}}^T\hat{\mathrm{F}}\right)^{-1}\hat{\mathrm{F}}^T\vec{\mathbf{Y}}$, полученные согласно (6.9), оказываются *смещенными*. Если число условных уравнений велико, а ошибки в факторах (элементы матрицы $\hat{\Delta}$) независимы между собой и имеют *диагональную* ковариационную матрицу

$$\hat{\mathrm{V}}_{\hat{\Delta}} = \left(0,\ \sigma^2_{x_1},\ \sigma^2_{x_2},\ \ldots,\ \sigma^2_{x_m}\right)\,, \tag{7.38}$$

то для величин смещения оценок параметров модели найдена формула

$$\hat{\mathrm{E}}\,\vec{\mathbf{B}} - \vec{\boldsymbol{\beta}} = N\left(\hat{\mathrm{F}}^T\hat{\mathrm{F}}\right)\breve{V}_{\hat{\Delta}}\vec{\mathbf{B}}\,, \tag{7.39}$$

где N — число условных уравнений; $\breve{V}_{\hat{\Delta}}$ — оценка матрицы (7.38). Смещение велико, если матрица $\hat{\mathrm{F}}^T\hat{\mathrm{F}}$ плохо обусловлена.

Приближенную несмещенную оценку погрешности вектора оценок параметров $\vec{\mathbf{B}}$ можно получить *в предположении малости погрешностей*, когда матрица $\hat{\Delta}$ близка к нулевой. Тогда

$$\breve{V}_{\vec{\mathbf{B}}} = s^2\left(\hat{\mathrm{F}}^T\hat{\mathrm{F}}\right)^{-1}, \tag{7.40}$$

и для s^2 имеем оценку

$$s^2 = \frac{1}{N-k}\left(\vec{\mathbf{Y}} - \hat{\mathrm{F}}\,\vec{\mathbf{B}}\right)^T\left(\vec{\mathbf{Y}} - \hat{\mathrm{F}}\,\vec{\mathbf{B}}\right)\,, \tag{7.41}$$

$k = m + 1$ — число оцениваемых параметров линейной модели.

При выводе выражения для смещения оценок параметров регрессионной модели при ненулевых ошибках факторов были сделаны достаточно жесткие предположения. В более общем случае задача превращается в трудноразрешимую нелинейную задачу. Одним из наиболее простых путей ее решения является получение смещений оценок параметров как функции дисперсий ошибок в отклике и факторов путем проведения обширных численных экспериментов, что будет рассмотрено в одной из следующих глав. Недостающие подробности рассмотренной ситуации можно найти в монографиях [23,24] списка литературы.

В качестве альтернативного метода улучшения можно посоветовать постепенное освобождение выборки от крайних значений невязок, что соответствует одному из робастных методов оценивания (см. следующую главу). Последнее позволяет искусственно уменьшить дисперсию невязок, что может уменьшить систематическое влияние случайных ошибок и уменьшить смещение оценок параметров регрессионной модели.

7.4. Выбор наилучшей модели регрессии

Вернемся вновь к линейной регрессионной задаче в классической постановке. Часто встречается ситуация, когда исследователь не имеет возможности заранее определить, какую регрессионную модель необходимо использовать, хотя имеется информация, что модель может быть линейной например. В этом случае возникает вопрос: какие факторы включать в модель для объяснения максимально большой доли дисперсии отклика? Желательно, чтобы количество включенных факторов

при этом было минимальным, чтобы не сделать процедуру оценивания и последующего использования модели для предсказания значений отклика слишком громоздкой. Очевидно, что качественная процедура выбора наилучшей регрессионной модели может снабдить исследователя информацией об изучаемом явлении, выявить факторы, реально связанные с явлением. Примером подобной задачи является используемое в фотографической астрометрии определение постоянных ПЗС-кадра или пластинки либо определение уравнений, связывающих координаты изображений на кадре с экваториальными координатами. Другим примером может служить задача об определении параметров звезд из показателей цвета многоцветной фотометрии, когда несколько показателей цвета могут быть в разной степени коррелированными, например, с абсолютной звездной величиной, эффективной температурой или показателем содержания металлов [Fe/H].

Существует несколько методов, служащих для решения поставленной задачи. **Метод оценивания значимости факторов**, включенных в регрессионную модель, с помощью рассмотренного в предыдущей главе частного коэффициента корреляции служит основой этих методов.

Сначала мы рассмотрим оригинальный метод, называемый **ступенчатым регрессионным методом**. Основная его идея заключается в следующем:

1. Сначала получается однофакторное уравнение регрессии, включающее фактор, наиболее коррелированный с откликом.
2. Затем находят невязки этого однофакторного уравнения регрессии $y_i - \check{y}_i$, эти невязки теперь рассматриваются как но-

вые значения отклика. Строится регрессия этого отклика на следующий фактор, наиболее коррелированный с новым откликом, хотя можно использовать и другие соображения для выбора.

Процесс продолжается до любой желаемой стадии. При этом получающееся уравнение не будет МНК-уравнением для включенных в него факторов. Оценки коэффициентов линейной модели, полученные ступенчатым методом, $\vec{\mathbf{B}}_C$, связаны с МНК-оценками через коэффициенты корреляции между факторами и по абсолютной величине меньше, чем МНК-оценки. Несмотря на то что этот метод всегда будет менее точным, то есть будет давать больший остаточный средний квадрат, чем метод наименьших квадратов, он имеет следующее преимущество: возможность управлять процедурой выбора факторов на основе экспертных оценок и априорной информации. Но настоящее МНК-уравнение обычно обладает лучшими свойствами в отношении предсказания значений отклика, чем уравнение, полученное ступенчатым методом. Впрочем, можно воспользоваться ступенчатым методом только для выбора значимых факторов, а затем для модели с выбранными факторами провести обычный МНК-анализ.

Перейдем к обзору методов, основанных на оценке значимости включаемых в модель факторов. В качестве наиболее очевидного, но и наиболее трудоемкого метода рассмотрим **метод всех возможных регрессий**. Он требует прежде всего построения каждого из всех мыслимых регрессионных уравнений, которые содержат свободный член x_0 и некоторое число факторов x_i. Пусть максимальное число мыслимых факторов равно r. Тогда мы должны рассматривать 2^r уравнений регрессии.

Например, если $r = 10$, нам придется рассмотреть 1 024 уравнения. Качество каждого из уравнений оценивается с помощью некоторого критерия. Таким критерием может быть, например, уже хорошо знакомый нам множественный коэффициент корреляции R^2, но можно использовать и сумму квадратов невязок s^2. В последнем случае исследованию уравнений помогает построение графика зависимости величины s^2 от количества включенных в модель факторов p.

Метод исключения более экономичен, чем метод всех возможных регрессий, поскольку в нем делается попытка исследовать только наилучшие регрессионные уравнения, содержащие определенное число факторов. Основные шаги этого метода сводятся к следующему:

1. Рассматривается регрессионное уравнение, включающее все мыслимые переменные.
2. Вычисляется величина частного F-критерия для каждого фактора в предположении, как будто бы она была последней переменной, введенной в регрессионное уравнение.
3. Наименьшая из полученных величина частного F-критерия, обозначаемая, допустим, F_L, сравнивается с заранее выбранным критическим значением F_0, в качестве которого часто выбирают значение около 3.5 или 4.0.
4. Если $F_L < F_0$, то анализируемый фактор x_L исключается из уравнения и производится перерасчет уравнения регрессии с включением оставшихся факторов.
5. Если $F_L > F_0$, то регрессионное уравнение оставляют таким, как оно было рассчитано.

Недостатком этого метода является то, что в случае плохой обусловленности матрицы $\hat{\mathrm{X}}^T\hat{\mathrm{X}}$ при большом числе включенных факторов уравнение может быть бессмысленным из-за ошибок округления.

Более распространено использование **шагового регрессионного метода**. Шаговый метод представляет собой попытку прийти к тем же результатам, что дает метод исключения. Но в этом методе действуют в обратном направлении, то есть включают факторы в модель до тех пор, пока уравнение не станет удовлетворительным. Порядок включения определяется с помощью частного коэффициента корреляции как меры важности переменных, еще не включенных в уравнение.

Основная процедура заключается в следующем. Прежде всего выбирается фактор, наиболее коррелированный с откликом (пусть это будет x_1), и находится регрессионное уравнение первого порядка. Затем проверяется, значима ли эта переменная. Если это не так, то мы должны согласиться с выводом, что наилучшая модель выражается уравнением $y = \langle y \rangle$. В противном случае необходимо найти второй фактор, который следует включить в модель.

Далее мы определяем частные коэффициенты корреляции для всех факторов, еще не включенных в уравнение. Теперь выбирается фактор, имеющий максимальный частный коэффициент корреляции, и находится второе регрессионное уравнение $\check{y} = f(x_1,\ x_2)$. Полное уравнение проверяется на значимость. Затем с помощью частного F-критерия проверяется значимость обоих факторов, а не только включенного последним, так как включение в уравнение второго фактора может изменить статус первого из-за обычно ненулевой коррелирован-

ности факторов между собой. Наименьшее из полученных значений частного F-критерия сравнивается с заданным критическим значением F_0, принимается решение об исключении соответствующего фактора. Такая проверка проводится на каждом шаге метода. В конечном счете процесс прекращается тогда, когда не удается ни исключить какой-либо из включенных факторов, ни включить новые.

Именно шаговый регрессионный метод можно рекомендовать как основной в тех случаях, когда нельзя из уже имеющейся информации надежно выбрать регрессионную модель заранее.

7.5. Нелинейный МНК

Определение 7.1. *Любая модель, вид которой не совпадает с известным нам уравнением для линейной по искомым параметрам модели, называется моделью нелинейной регрессии и может быть записана в виде*

$$y_i = f\left(x_{1i},\ \ldots,\ x_{pi};\ \beta_1,\ \ldots,\ \beta_m\right) + e_i, \tag{7.42}$$

$f(\cdot)$ — нелинейная функция параметров $\beta_1,\ \ldots,\ \beta_m$; x_{ik} — факторы модели; y_i — измеряемые отклики; e_i — некоррелированные ошибки.

Примером истинно нелинейной функции может служить функция

$$f\left(x_i;\ \beta_1,\ \beta_2\right) = \beta_1 + \exp\left(\beta_2\, x\right)\ . \tag{7.43}$$

Если модель линейна, то МНК-оценки параметров будут оптимальными, так как они являются *несмещенными оценками с минимальной дисперсией*. Но **для нелинейных моделей универсальные аналитические методы получения оптимальных оценок параметров отсутствуют**. Также для нелинейных моделей не гарантируется наличие или единственность решения даже для невырожденной матрицы системы уравнений.

Рассмотрим нелинейную модель в общем виде:

$$\begin{aligned} y &= f\left(x_1,\ x_2,\ \ldots,\ x_k;\ \beta_1,\ \beta_2,\ \ldots,\ \beta_m\right) + \epsilon = \\ &= f\left(\vec{\mathbf{X}},\ \vec{\boldsymbol{\beta}}\right) + \epsilon\,. \end{aligned} \tag{7.44}$$

Пусть имеем n измерений факторов и отклика, тогда сумма квадратов остаточных отклонений есть

$$SS_0 = \sum_{i=1}^{n} \left(y_i - f\left(x_{i1},\ \ldots,\ x_{ik};\ \beta_1,\ \ldots,\ \beta_m\right)\right)^2 . \tag{7.45}$$

Нормальные уравнения можно получить обычным способом, то есть дифференцированием SS_0 как функции от параметров по параметрам β_i и приравниванием частных производных нулю. Для модели (7.43), например, получаем систему из двух уравнений

$$\begin{aligned} &\sum_{i=1}^{n} \left(y_i - b_1 - \exp\left(b_2\, x_i\right)\right) = 0\,, \\ &\sum_{i=1}^{n} x_i \exp\left(b_2 x_i\right) \left(y_i - b_1 - \exp\left(b_2\, x_i\right)\right) = 0\,, \end{aligned} \tag{7.46}$$

или, раскрывая скобки,

$$\begin{aligned} \sum_{i=1}^{n} y_i &= b_1 n + \sum_{i=1}^{n} \exp\left(b_2\, x_i\right) \,, \\ \sum_{i=1}^{n} x_i \exp\left(b_2 x_i\right)\, y_i &= b_1 \sum_{i=1}^{n} x_i \exp\left(b_2 x_i\right) + \\ &+ \sum_{i=1}^{n} x_i \exp\left(2\, b_2\, x_i\right) \,, \end{aligned} \tag{7.47}$$

где b_1 и b_2 — выборочные оценки параметров β_1 и β_1.

Приведенный пример показывает, что в случае нелинейных условных уравнений получающиеся нормальные уравнения трудноразрешимы. Общего метода решения их не существует. Поэтому в нелинейном регрессионном анализе применяют почти исключительно численные методы поиска минимума остаточной суммы квадратов как функции m неизвестных параметров модели. Однако в некоторых, не очень редких случаях срабатывает **метод последовательных приближений**, методику применения которого покажем на примере системы (7.47).

Если у нас есть хорошая предварительная оценка одного из параметров, например b_1, можно решить первое из уравнений системы (7.47) относительно b_2, например, *методом Ньютона*. Затем можно использовать второе уравнение системы, подставив в него полученную оценку b_2 и решая его относительно b_1, получая, таким образом, следующее приближение для b_1. Процесс, если он сходится, продолжают до того момента, пока изменение на шаге процесса в оценках искомых параметров не станет меньше выбранной достаточно малой величины.

Если нулевые приближения для параметров известны плохо, можно попытаться использовать **метод парабол**, который заключается в следующем. Зафиксируем значение одного параметра, например b_1, и вычислим квадраты разностей между левой и правой частями первого уравнения для трех значений b_2. Проведем параболу через три точки, задаваемые величинами b_2 и квадратами разностей, найдем положение минимума зависимости, который можно взять в качестве следующего приближения к значению b_2. Используя это значение, из второго уравнение таким же способом находим следующее приближение для b_1 и т. д. до тех пор, пока сумма квадратов отклонений не станет изменяться очень мало.

Рассмотрим идеи других численных методов решения задач минимизации целевой функции типа уравнения (7.45). Наиболее известным из таких методов является **метод линеаризации**, когда решение нелинейной задачи заменяется решением ряда задач линейной регрессии. Пусть тем или иным способом получено нулевое приближение вектора искомых параметров $\vec{\mathbf{B}}_0 = (b_{01},\ b_{02},\ \ldots,\ b_{0m})$. Разлагая $f\left(\vec{\mathbf{X}},\ \vec{\mathbf{B}}\right)$ в степенной ряд в окрестностях точки $\vec{\mathbf{B}}_0$ и ограничиваясь первыми степенями разностей $\vec{\mathbf{B}} - \vec{\mathbf{B}}_0$, получим линейное относительно этой разности соотношение

$$y = f\left(\vec{\mathbf{X}},\ \vec{\mathbf{B}}_0\right) + \vec{\mathbf{D}}\left(\vec{\mathbf{X}},\ \vec{\mathbf{B}}_0\right)\left(\vec{\mathbf{B}} - \vec{\mathbf{B}}_0\right), \qquad (7.48)$$

где $\vec{\mathbf{D}}\left(\vec{\mathbf{X}},\ \vec{\mathbf{B}}_0\right)$ — вектор частных производных от функции $f\left(\vec{\mathbf{X}},\ \vec{\mathbf{B}}\right)$ по параметрам, взятым в точке $\vec{\mathbf{B}}_0$. Теперь с помощью линейного МНК можно найти оценки разностей $\vec{\mathbf{B}} - \vec{\mathbf{B}}_0$

и первое приближение $\vec{\mathbf{B}}_1 = \vec{\mathbf{B}}_0 + \left(\vec{\mathbf{B}} - \vec{\mathbf{B}}_0\right)$. Аналогично ищутся следующие приближения. Процесс продолжается до достижения выбранной точности приближения. Для контроля желательно на каждом шаге вычислять остаточную сумму квадратов SS_0; если эта величина начнет прогрессивно увеличиваться или испытывать сильные колебания, следует выбрать иные значения для нулевого приближения и вновь повторить вычисления.

В настоящее время метод линеаризации применяется редко, чаще используют **градиентные методы**, **симплекс-метод** и **метод Марквардта**, описание которых можно найти, например, в книге Й. Барда [25]. Метод Марквардта и другие методы нелинейного оценивания параметров можно найти в пакетах математических и статистических вычислений.

Имеется еще один метод, для которого легко написать программу и который часто использовался при большом числе параметров в модели. Это **метод случайного поиска**. Идея метода очень проста. Для каждого параметра задается интервал, в котором точное значение параметра должно содержаться почти наверняка. Затем с помощью датчика случайных чисел, подходящим образом распределенных, задаются случайные значения для каждого параметра и вычисляются остаточные суммы квадратов. При большом числе использованных случайных точек мы можем надеяться на хороший подбор оценок искомых параметров. Ясно, что для надежного оценивания параметров нелинейной регрессионной модели требуется очень большое число точек, однако, если модель сложна для вычислений, этот метод при большом количестве параметров может быть достаточно экономичен.

В случае нелинейной по параметрам регрессии оценить точность результата сложнее, чем в линейном случае [26]. В качестве очевидного и простого способа может служить разбиение выборки на 3—5 частей-подвыборок и дальнейшее оценивание параметров отдельно по каждой из подвыборок с последующим усреднением полученных оценок и вычислением выборочных дисперсий, но для надежных оценок должны использоваться более строгие методы [26].

8. РОБАСТНОЕ ОЦЕНИВАНИЕ

8.1. Робастное оценивание параметров выборочных распределений

Определение 8.1. ***Робастными или устойчивыми методами оценивания*** *называются такие методы, при которых значения реализаций случайных величин, находящиеся в крыльях выборочного распределения, либо попадание в выборку некоторого малого количества элементов из другой генеральной совокупности мало влияют на оценивание параметров распределения.*

Робастные методы оценивания применяются в случаях, когда в выборке могут появиться грубые промахи, что эквивалентно присутствию в выборке элементов из другого распределения, характеризуемого значительно большей дисперсией. В этом случае на оценках параметров распределения сказывается появление даже небольшого числа таких мешающих значений. Подобный эффект и называется **неустойчивостью оценок**. Особенно сильно он проявляется при множественной регрессии, в этом же случае он наиболее опасен, так как влияние многих факторов может замаскировать проявление неустойчи-

вости. В конце главы разобран пример, когда мешающие значения не являются грубыми промахами и их появление обусловлено особенностями задачи, но такие значения резко уменьшают устойчивость оценки математического ожидания. Все робастные методы тем или иным способом подавляют влияние на оценивание значений, попадающих в крылья распределений.

Рассмотрим некоторые методики получения робастных оценок.

Винзоризованные оценки (иногда встречается название «оценки Винзора») применяются при оценивании среднего и дисперсии распределений, при построении доверительных интервалов и при проверке гипотез относительно генерального среднего. В этой процедуре крайние значения вариационного ряда не отбрасываются, а изменяются.

Определение 8.2. *Обозначим через* $y_1 \leqslant y_2 \leqslant \ldots \leqslant y_n$ *упорядоченный ряд для* n *наблюдений —* ***вариационный ряд****. Тогда* g***-винзоризованные наблюдения*** *получаются заменой первых* g *наблюдений на* y_{g+1}*, а* g *последних — на* y_{n-g}*, при этом* $1 \leqslant g < n/2$*.*

Таким образом, по определению мы переходим к переменным z_i, задаваемым по правилу

$$\begin{aligned} z_1 = z_2 = \ldots = z_g = y_{g+1}\,, \\ z_{g+i} = y_{g+1} \text{ для } 2 \leqslant i \leqslant n-2g-1\,, \\ z_n = z_{n-1} = \ldots = z_{n-g+1} = y_{n-g}\,. \end{aligned} \tag{8.1}$$

При этом оценками среднего и дисперсии исходного распределения служат соответственно величины

$$\langle z \rangle = \frac{1}{n} \sum_{i=1}^{n} z_i \,, \tag{8.2}$$

и

$$s_z^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(z_i - \langle z \rangle \right)^2 . \tag{8.3}$$

Приближенный доверительный интервал таких оценок для уровня значимости α задается выражением

$$\langle z \rangle \pm \left(\frac{n-1}{h-1} \right) \left(\frac{s_z}{\sqrt{n}} \right) t_{1-(\alpha/2)} \, (h-1), \tag{8.4}$$

где $h = n - 2g$; $t_{1-(\alpha/2)} \, (h-1)$ — коэффициент Стьюдента для доверительной вероятности $1 - (\alpha/2)$ и $h-1$ степеней свободы.

Для проверки гипотезы $H_0 : \xi = \xi_0$, где ξ_0 — предполагаемое значение математического ожидания для ряда $\{z_i\}$, соответствующий g-винзоризованный односторонний t-критерий использует статистику

$$t = \frac{(h-1)\sqrt{n} \, (\langle z \rangle - \xi)}{(n-1) \, s_z} \tag{8.5}$$

с $h - 1 = n - 2g - 1$ степенями свободы.

Строгого метода выбора значений g нет. Его величина подбирается исходя из предположений о виде функции распределения выборки. Желательно использовать минимальные значения g.

Можно не заменять крайние значения вариационного ряда, а совсем выбросить их из рассмотрения, при этом для достаточно больших выборок получающееся среднее распределено приблизительно нормально.

Еще одним способом уменьшения влияния крайних членов выборки является введение низких весов для величин, попадающих в крылья распределения. Веса можно вводить различными способами. К таким оценкам относятся так называемые M-**оценки Хубера**. Рассмотрим идею получения M-оценок на примере линейной регрессионной задачи. Оценки параметров регрессионной модели $\vec{\mathbf{B}}_M$ получаются из решения минимизационной задачи

$$\sum_{k=1}^{n} \rho(e_k) = \sum_{k=1}^{n} \rho\left(y_k - \sum_{i=1}^{n} b_k\, x_{ik}\right) = \min\,, \tag{8.6}$$

где $\rho(z)$ — некоторая функция, которая и определяет свойства оценок $\vec{\mathbf{B}}_M$. В зависимости от способа выбора функции $\rho(z)$ возникают различные виды оценок. Когда эта функция выпуклая, задача (8.6) имеет единственное решение. Таким образом, идея получения M-оценок Хубера есть уменьшение влияния далеко отстоящих от «идеальной» регрессионной прямой точек.

Имеются способы выбора функции $\rho(z)$ в зависимости от искажения распределения невязок — «интенсивности» грубых ошибок [24]. Численные исследования свидетельствуют, что M-оценки в ряде случаев имеют значительно более высокую эффективность, чем МНК, даже при малых интенсивностях «загрязнения» выборки. При этом моделирование показывает, что

и при несимметричном распределении «загрязнения» смещение оценок параметров регрессионной модели пренебрежимо мало.

Важным методом получения устойчивых оценок параметров распределений является **оценивание не среднего значения, а моды распределения**. Так как положение максимума (мода) выборочного распределения определяется большим количеством членов выборки с малыми отклонениями от моды, а крылья распределения не влияют на положение моды, то обеспечивается хорошая устойчивость оценки этой величины при достаточно больших выборках. Использование M-оценок Хубера, учитывая заложенную в этот метод идеологию, позволяет эффективно оценивать моду распределения. Ошибку положения моды можно оценить, разбив выборку на несколько непересекающихся подвыборок и определив оценки моды для каждой из подвыборок в отдельности. Затем из этих оценок можно получить среднее и выборочную дисперсию, учитывая при этом уменьшение объема для подвыборок.

Известно, что *приблизиться к оценке моды можно, переходя от минимизации суммы квадратов уклонений к минимизации сумм меньших степеней модулей уклонений*, что и делается с помощью M-оценок Хубера. Такую методику можно применить в регрессионном анализе, если известно, что выборка загрязнена резко выделяющимися значениями (промахами), или если крылья распределения невязок усилены. При этом задача становится нелинейной даже для линейной по параметрам модели, однако усложнение задачи в определенных случаях окупается получением более устойчивых оценок параметров модели. Недостатками этой методики являются уменьше-

ние «резкости» главного, глобального минимума при уменьшении степени и возможность появления локальных минимумов, особенно при небольших выборках.

8.2. Практическое применение робастного оценивания на примере определения частоты вращения диска Галактики

Приведем практический пример, который основывается на работе, выполненной А. В. Локтиным совместно с Г. В. Бешеновым.

Пусть мы по данным о движениях рассеянных звездных скоплений определяем частоту вращения диска Галактики на круге Солнца. Обычно частоту вращения диска Галактики ω_0 на круге Солнца, где расстояние от оси вращения Галактики R равно расстоянию Солнца от этой оси R_0, оценивается как разность между постоянными Оорта: $\omega_0 = A - B$. Однако можно обойтись без разложения поля круговых скоростей, не получая оценки постоянных Оорта. Для этого используются формулы Боттлингера, описывающие поле круговых скоростей в диске Галактики:

$$v_r = R_0 \left(\omega(R) - \omega_0\right) \sin l \cos b \,, \tag{8.7}$$

$$v_t = R_0 \left(\omega(R) - \omega_0\right) \cos l \cos b - \omega(R)\, r \cos b \,, \tag{8.8}$$

где v_r и $v_t = 4.738\, r\, \mu_l$ — лучевая скорость и тангенциальная скорость объектов, исправленные за движение Солнца в пространстве; l и b — галактические долгота и широта. Исключая

из (8.7), (8.8) функцию $\omega(R)$ и разрешая полученное выражение относительно ω_0, получаем выражение

$$\omega_0 = \omega(R_0) = \frac{v_r\,(R_0 \cos l - r \cos b) - v_t\,R_0 \sin l \cos b}{R_0\, r \sin l \cos^2 b}\,, \qquad (8.9)$$

которое и было использовано для оценивания величины ω_0 по движениям 165 рассеянных звездных скоплений. Отметим, что для галактических долгот около 0 и 180° знаменатель выражения (8.9) близок к нулю, так что малые ошибки в наблюдаемых величинах ведут к большим изменениям оценок ω_0. Поэтому необходимо удалить из выборки объекты, расположенные вблизи центра и антицентра Галактики, а для оставшихся ввести веса, обратно пропорциональные величинам знаменателя выражения (8.9). Так как наблюдаемые величины скоростей и расстояний до скоплений определяются с ошибками, то необходимо также ввести веса, пропорциональные качеству наблюдательных данных.

Гистограмма распределения полученных оценок ω_0 приведена на рис. 8.1. Отметим, что эта гистограмма построена не подсчетом чисел попадания оценок в интервалы аргумента, а суммированием весов, так что ее можно назвать **гистограммой накопленных весов**. Для неравноточных данных это уменьшает вклад оценок, полученных по низкоточным исходным данным. В этом случае, если мы получаем оценки параметров распределения по значениям гистограммы, в результате вычисляем взвешенные оценки.

На рис. 8.1 хорошо выделяется центральный максимум, но видно также, что крылья распределения очень сильны. Крылья образованы как оценками, полученными по движе-

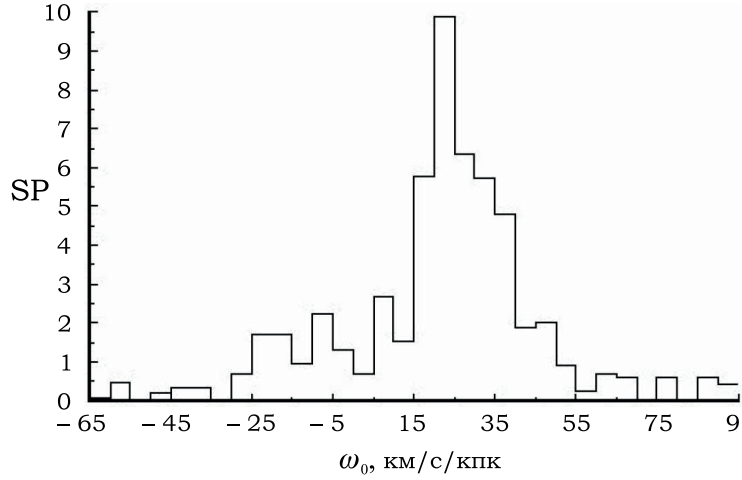

Рис. 8.1. Гистограмма распределения оценок частоты вращения диска Галактики на круге Солнца ω_0

ниям объектов с большими ошибками в данных наблюдений, так и объектов, для которых знаменатель в выражении (8.9) мал, при этом видно, что крылья являются асимметричными. Очевидно, что если мы будет вычислять среднее значение $\langle\omega_0\rangle$, то его величина в основном будет определяться распределением удаленных от максимума распределения индивидуальными оценками, то есть среднее при усиленных крыльях распределения является оценкой неустойчивой. Поэтому мы решили определить моду распределения, вычисляя эту оценку как средневзвешенное для пяти интервалов гистограммы, симметрично расположенных относительно максимума. Это дало оценку $\omega_0 = 25.6$ км/с/кпк. Погрешность данной величины определили, разделив выборку на три непересекающиеся подвыборки. Для каждой подвыборки были получены оценки

моды распределения и выборочная дисперсия s_1^2 трех оценок. Так как подвыборки имеют в три раза меньший объем, чем вся выборка, то дисперсия среднего для оценки по всей выборке есть $s_0^2 = s_1^2/3$. Далее можно обычным способом оценить доверительный интервал.

Распределения с усиленными крыльями часто встречаются в звездной астрономии, поэтому к робастному оцениванию следует относиться с большим вниманием.

9. СЛУЧАЙНЫЕ ПРОЦЕССЫ

9.1. Характеристики случайных процессов

Определение 9.1. *Временны́м рядом обычно называют упорядоченное во времени множество наблюдений над некоторой случайной величиной на конечном интервале времени. Временной ряд при этом может быть дискретным и непрерывным.*

Примером *непрерывного ряда* является **регистрограмма**, то есть непрерывная запись образа некоторого процесса, протекающего во времени (например, изменение блеска переменной звезды). Однако при обработке наблюдений с регистрограммы снимают отдельные отсчеты, так что и в этом случае мы имеем дело с *дискретным рядом*. К временным рядам можно отнести также выборки, упорядоченные по пространственной или иной координате, например, запись при сканировании спектра щелью фотометра, график погрешности микрометренного винта и т. д.

С точки зрения приведенного выше общего определения *любая выборка* из некоторой генеральной совокупности тоже *может рассматриваться как временной ряд*, поскольку ее элементы получают значения в эксперименте не одновременно, так что их всегда можно упорядочить по времени. Однако

в случае обычной некоррелированной выборки моменты времени не фиксируются, так как если распределение генеральной совокупности не меняется, то фиксация времени ничего не прибавляет к информации о случайной величине.

Определение 9.2. ***Случайный процесс*** *— это процесс, когда со временем изменяется закон распределения, притом случайно, а элементы выборки коррелированы. В отличие от временного ряда случайный процесс задается на всей числовой оси и является обычно непрерывной функцией времени.*

Выделяются много классов случайных процессов, которые характеризуются самыми различными способами. Рассмотрение измеряемых в опыте величин как реализаций некоторой случайной величины позволяет вводить для случайного процесса те же характеристики, что и для случайных величин, но так как *случайный процесс — это не случайная величина*, то вводятся некоторые новые характеристики. Примерами случайного процесса могут служить кривые блеска квазаров или неправильных переменных звезд, случайные флуктуации заряда или напряжения светоприемников и т. д.

Определение 9.3. *Основными характеристиками случайного процесса $X(t)$ являются его математическое ожидание $\hat{\mathrm{E}}\left(X(t)\right) = \xi(t)$ и дисперсия $\hat{\mathrm{E}}\left(X(t) - \xi(t)\right)^2 = \sigma^2(t)$, а также* ***ковариационная функция****, представляющая математическое ожидание произведений отклонений случайной функции от тренда в моменты t и $t' = t + \tau$ при сдвиге τ,*

$$C(t,\ t') = \hat{\mathrm{E}}\left(\left(X(t) - \xi(t)\right)\left(X(t') - \xi(t')\right)\right), \qquad (9.1)$$

где t и t' пробегают все возможные значения аргумента t.

При фиксированных значениях t и t' значение ковариационной функции есть ковариация между случайными величинами $X(t)$ и $X(t')$. Отметим, что все три характеристики — $\xi(t)$, $\sigma^2(t)$ и $C(t,\ t')$ — *являются неслучайными функциями своих аргументов.*

Определение 9.4. *Если некоторый процесс $Y(t)$ представляет собой сумму случайного процесса $X(t)$ и неслучайной функции $\phi(t)$, то*

$$\hat{\mathrm{E}}\left(Y(t)\right) = \xi(t) + \phi(t), \qquad (9.2)$$

а дисперсия и ковариационная функция будут такими же, как у случайного процесса $X(t)$.

От ковариационной функции легко перейти к **корреляционной функции**

$$\rho(t,\ t') = \frac{C(t,\ t')}{\sqrt{\sigma^2(t)\,\sigma^2(t')}}, \qquad (9.3)$$

которая дает значения **коэффициентов корреляции** между $X(t)$ и $X(t')$.

Очевидно, что *в общем случае оценить параметры случайного процесса невозможно*, так как невозможно получить несколько значений случайной величины $X(t)$ в один и тот же момент времени. Поэтому оценку параметров проводят только для некоторых классов случайных процессов, в частности, *для стационарных случайных процессов*, которые и будут пред-

метом нашего дальнейшего рассмотрения. Сказанное легко понять, рассматривая рис. 9.1, на котором показан пример некоторого случайного процесса. При каждом значении аргумента t мы можем вычислить значение, например, дисперсии в том случае, если сможем наблюдать процесс очень большое (бесконечное) количество раз, — мы должны рассматривать множество реализаций случайного процесса, чтобы для каждого значения t была возможность оценить параметры процесса. На практике это можно сделать, например, в случае, если на график мы нанесем много отрезков случайного процесса, *отождествляя* для них отрезки аргумента. При этом можно рассматривать функцию и плотность распределения случайного процесса, а математическое ожидание и дисперсию рассматривать как параметры данных распределений. Но для этого параметры процесса не должны меняться со временем.

Определение 9.5 (Строгая стационарность)**.** *Случайный процесс называется* ***стационарным*** *в том случае, если функция и плотность распределения случайного процесса и, как следствие, математическое ожидание и дисперсия зависят только от разности значений аргумента процесса (времени)* $\Delta t = t - t'$*, но не зависят от самого значения аргумента* t*.*

Отметим еще следующее. Зачастую случайный процесс может рассматриваться как непрерывная функция времени. Однако при обработке мы выбираем некоторое дискретное подмножество значений аргумента и ведем обработку дискретного сигнала. Фактически мы заменяем непрерывную функцию ступенчатой при обработке выборки из «генеральной совокупности». Такой переход от непрерывной функции называется

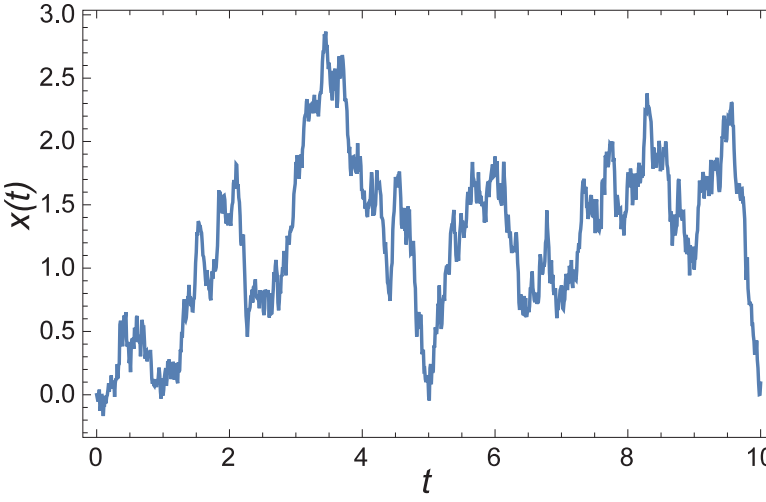

Рис. 9.1. Пример случайного процесса

дискретизацией. При этом возникает проблема: какой следует взять интервал дискретизации Δt, чтобы потерять минимум информации, содержащейся в исходном сигнале? Пусть мы выбрали шаг Δt. Рассмотрим некоторую синусоиду и отметим на ней точки, соответствующие значениям аргумента $t_0 + i\,\Delta t$, где i — номер точки, считая слева. Если точки распределены по графику синусоиды редко, то может найтись синусоида меньшей частоты, проходящая через эти точки. Такое положение нас не устраивает, поскольку это ложный сигнал, появившийся вследствие процесса дискретизации и не содержащийся в исходном временном ряду. Чтобы при анализе не возникали ложные частоты, следует шаг дискретизации выбирать достаточно малым.

Решение проблемы выбора шага дискретизации дается *теоремой Котельникова*, которая лежит в основе всех импульсных систем связи [27–29].

Теорема 9.1 (Котельникова—Найквиста—Шеннона)**.** *Любую функцию $F(t)$, состоящую из частот от 0 до f_c, можно непрерывно передавать с любой точностью при помощи чисел, следующих друг за другом через $1/(2f_c)$ с. Частота f_c, равная половине частоты дискретизации, называется* ***частотой Найквиста****.*

Следствием из этой теоремы является то, что аналоговый сигнал может быть восстановлен с какой угодно точностью по своим дискретным отсчетам, взятым с частотой $f > 2f_c$, где f_c — максимальная частота, присутствующая в спектре реального сигнала. Для реальных сигналов, а не в идеальных условиях теоремы статистически достоверное восстановление аналогового сигнала требует частоты выборки $f > 5f_c$.

9.2. Оценивание параметров стационарного случайного процесса

Определение 9.6 (Слабая стационарность)**.** *Стационарным в узком смысле называется случайный процесс, при котором закон распределения случайной величины, а также ее математическое ожидание и дисперсия не изменяются со временем, а ковариационная и корреляционная функции непрерывны и зависят только от одной переменной $\tau = t - t'$, то есть стационарный случайный процесс является* ***инвариантным*** *относительно изменения начала отсчета.*

Пусть $x(t_0),\ x(t_1),\ \ldots,\ x(t_n)$ — временной ряд, представляющий собой реализацию некоторого случайного процесса $X(t)$, причем временные интервалы Δt между соседними значениями аргумента одинаковы, что является обязательным для оценивания ковариационной функции. Таким образом построенные значения называют **эквидистантными**. Поскольку приращения аргумента постоянны, мы далее будем обозначать элементы временного ряда как $x_0,\ x_1,\ \ldots,\ x_n$, а иногда номер значения использовать в качестве аргумента.

На практике часто бывает так, что временной ряд включает неслучайную составляющую — так называемый **временной тренд** (непериодическую составляющую), либо периодическую составляющую, либо ту и другую вместе. Перед исследованием случайного процесса такие тренды *необходимо устранить*. Для исключения тренда, то есть когда $\hat{\mathrm{E}}\,(X(t)) = \xi(t) \neq 0$, приходится прибегать к регрессионному анализу — отысканию параметров неслучайной функции $\xi(t)$ с последующим вычитанием из каждого значения x_i поправки $\check{\xi}(t_i)$. В качестве модели для неслучайной функции обычно используют полиномы невысокой степени, экспоненты, а для исключения периодической составляющей — тригонометрические полиномы, в том числе отрезки тригонометрических рядов.

Оценкой математического ожидания для стационарного временного ряда является **среднее по множеству измерений**

$$\check{\xi} = \langle x \rangle = \frac{1}{n+1} \sum_{i=0}^{n} x_i \,. \tag{9.4}$$

Если было правильно проведено исключение тренда, то *среднее будет близко к нулю*, так как вместе с трендом автоматически исключается и постоянная составляющая случайного процесса, однако это не влияет на оценки дисперсии и ковариационной функции.

Дисперсию временного ряда можно оценить по привычной формуле

$$\breve{\sigma}_0^2 = s^2 = \frac{1}{n} \sum_{i=0}^{n} \left(x_i - \langle x \rangle\right)^2 \ . \tag{9.5}$$

Рассмотрим теперь *оценку ковариации* i-го элемента временного ряда с $(i+k)$-м элементом, то есть оценку *ковариационной функции* $C(k)$, так как для стационарного случайного процесса эта функция зависит только от величины временного сдвига (лага), равного в данном случае $k\,\Delta t$, а величину Δt можно выбором единиц измерения сделать равной единице. Используя обычную формулу для выборочной ковариации, получаем

$$\breve{C}(k) = \frac{1}{n-k} \sum_{i=0}^{n-k} \left(x_i - \langle x \rangle\right) \left(x_{i+k} - \langle x \rangle\right) \ , \tag{9.6}$$

$$\breve{C}(0) = \breve{\sigma}_0^2 = s^2. \tag{9.7}$$

Вычисляемые по этой формуле значения $\breve{C}(k)$ имеют различные числа степеней свободы, равные $n-k$, следовательно, и различную точность. Однако если n достаточно велико, то, пожертвовав частью информации, можно получить значения $\breve{C}(k)$ с постоянным числом степеней свободы $K < n$, ограничив временной лаг значением $\tau = K\,\Delta t$. Формулой для оцен-

ки ковариационной функции в этом случае будет

$$\breve{C}(k) = \frac{1}{K} \sum_{i=0}^{K} (x_i - \langle x \rangle)\ (x_{i+k} - \langle x \rangle)\ . \tag{9.8}$$

Необходимо заметить, что поскольку $i + k \leqslant n$ и, следовательно, $K + k \leqslant n$, то для получения K одинаковых по точности оценок $\breve{C}(k)$ необходим временной ряд, содержащий $2\,K$ элементов.

Поскольку дисперсия у всех отрезков стационарного временного ряда одна и та же, то для перехода к корреляционной функции $\breve{R}(k)$ достаточно разделить каждое значение $\breve{C}(k)$ на оценку дисперсии s^2, которая, как видно из сравнения формул (9.5) и (9.6), равна $\breve{C}(0)$.

Отметим, что *если стационарный случайный процесс содержит не исключенную периодическую составляющую* с периодом T, то и ковариационная и корреляционная функции будут содержать периодические составляющие с периодом $T/\Delta t$, на чем основано применение **автокорреляционной функции** для анализа временного ряда на существование периодической составляющей. Для временного ряда как непрерывной функции автокорреляционная функция записывается в виде интеграла

$$R_{ff}(\tau) = \frac{1}{T} \int_0^T f(t)\, f(t+\tau)\, dt\,, \tag{9.9}$$

так что *автокорреляционная функция представляет свертку временного ряда самого с собой*; двойной индекс $f\,f$ обозначает, что вычисляется корреляция функции f с самой собой. Значимое отличие функции $R_{ff}(\tau)$ от нуля для выбранного значения

сдвига τ говорит о возможной корреляции между собой элементов временного ряда, сдвинутых по отношению друг к другу на величину τ.

Для сравнения двух случайных процессов $f(t)$ и $g(t)$ можно использовать **функцию взаимной корреляции**

$$R_{fg}(\tau) = \lim_{T\to\infty} \int_0^T f(t)\, g(t+\tau)\, dt\,, \tag{9.10}$$

откуда легко записать формулу для дискретного случая, аналогичную (9.8).

9.3. Оценка спектральной плотности стационарного случайного процесса

Ковариационная (корреляционная) функция не очень удобна для интерпретации. В частности, синусоидальный сигнал распределен по всем значениям ее аргумента. Поэтому перейдем *от временной терминологии к частотной*. Этому соответствует, при анализе некоторого временного ряда, переход от ковариационной либо корреляционной функции к так называемой **функции спектральной плотности** $s(\omega)$, являющейся Фурье-преобразованием от $C(\tau)$ либо $R(\tau)$.

Известно, что любой стационарный случайный процесс можно представить в виде некоторой суммы конечного или бесконечного числа гармонических составляющих, при этом *спектральная мощность каждой гармоники пропорциональна квадрату ее амплитуды, а значит и ее дисперсии.*

Определение 9.7. *Функция спектральной плотности показывает распределение дисперсии временного ряда по частотам преобразования Фурье.*

Функцию спектральной плотности $s(\omega)$ определим как преобразование Фурье от корреляционной функции $R(\tau)$ и с учетом того, что $R(\tau)$ — четная функция:

$$s(\omega) = \int\limits_{-\infty}^{+\infty} R(\tau) \exp(-i\omega\tau)\, d\tau = 2 \int\limits_{0}^{+\infty} R(\tau)\, \cos(\omega\tau)\, d\tau\,. \quad (9.11)$$

Спектральная плотность есть четная функция частоты. При известной спектральной плотности обратное преобразование Фурье дает выражение для корреляционной функции через интеграл (или сумму) от функции спектральной плотности

$$R(\tau) = \frac{1}{2\pi} \int\limits_{0}^{+\infty} s(\omega)\, \cos(\omega\tau)\, d\omega\,. \quad (9.12)$$

Используя выражения (9.12), получим связь между дисперсией случайного процесса D и его спектральной плотностью:

$$D \equiv \sigma_0^2 = R(0) = \frac{1}{2\pi} \int\limits_{0}^{+\infty} s(\omega)\, d\omega\,. \quad (9.13)$$

Обозначив спектральную плотность (*относительную мощность*), приходящуюся на круговую частоту ω_i, символом $\sigma^2(\omega_i)$, можем записать выражение для **дисперсии ста-**

ционарного случайного процесса

$$D = \sum_{j=0}^{\infty} \sigma^2(\omega_j)\,. \tag{9.14}$$

В непрерывном случае суммирование в (9.14) следует заменить интегралом, а $\sigma^2(\omega_j)$ — заменить на $\sigma^2(\omega)\,d\omega$, где $\sigma^2(\omega)$ носит название **спектральной плотности**, то есть представляет собой *мощность, приходящуюся на единичный интервал частот.*

Оценки спектральной плотности стационарного случайного процесса в шкале обычных частот $\nu_i = \omega_i/2\pi$ могут быть получены с помощью формулы

$$\breve{\sigma}^2\,(\nu_j) = 2\left(\breve{C}(0) + 2\sum_{k=1}^{K-1} \breve{C}(k)\cos\left(2\,\pi k \nu_j\right)\right), \qquad j = 1,\ \ldots,\ K, \tag{9.15}$$

причем, в силу дискретности как временного ряда, так и шкалы частот, функцию $\breve{\sigma}^2\,(\nu_j)$ *следует рассматривать как гистограмму*: каждое ее значение представляет собой оценку средней спектральной плотности на интервале $\nu_j \pm \Delta\nu/2$, где $\Delta\nu = \nu_{j+1} - \nu_j$ — ширина интервала гистограммы. При выборе значений частот следует учитывать, что выборочная ковариационная функция имеет K значений, то есть при $\Delta t = 1$ самой низкой может быть частота $\nu_1 = 1/K$ (период равен K), а самой высокой $\nu_K = 1$ (период равен единице). Обычно частоты выбирают кратными низшей частоте ν_1, и тогда мы имеем K значений частоты для функции спектральной плотности.

Выборочную оценку спектральной плотности обычно называют **периодограммой**, чтобы отличить ее от спектра мощности, характеризующей случайный процесс, а не временной ряд, построенный по реализации случайного процесса.

Выборочная оценка спектральной плотности (периодограмма) $\breve{\sigma}^2(\nu_j)$ характеризует свойства временного ряда, построенного по реализации случайного процесса. Временной ряд строится по реализации случайного процесса умножением конкретной реализации случайного процесса на прямоугольную функцию с шириной, равной интервалу времени измерений, и на **функцию-гребенку**, что соответствует процессу дискретизации. Эти преобразования приводят к тому, что спектр мощности отличается от периодограммы на функцию, убывающую при увеличении интервала времени измерений и увеличении числа точек дискретизации.

Оценка спектральной плотности (9.15) *не является состоятельной* и, как следствие, *несмещенной*. При этом точность значений не увеличивается с увеличением выборки. На практике обычно используют более общее выражение для спектральной плотности

$$\breve{\sigma}^2(\nu_i) = 2\left(1 + 2\sum_{k=1}^{F-1} \breve{C}(k)\, W(k) \cos\left(\frac{\pi\, i\, k}{F}\right)\right), \qquad i = 1,\ \ldots,\ F, \tag{9.16}$$

где $W(k)$ — весовая функция или **окно**; коэффициенты $\breve{C}(k)$ — значения корреляционной функции; F — величина, равная половине периода, по которому корреляционная функция разлагается в ряд Фурье.

В качестве весовой функции окна можно, например, использовать окно Бартлетта

$$W(k) = \begin{cases} 1 - \dfrac{k}{L} & \text{при } k \leqslant L, \\ 0 & \text{для других } k, \end{cases} \tag{9.17}$$

где L — **точка отсечения корреляционной функции** (см. ниже). Введение окна аналогично сканированию изображения спектра звезды щелью и является сглаживающей процедурой. Существует немало других видов весовых функций, отличных от (9.17). Выборочные оценки спектральной плотности, получаемые с применением окна (9.17) и некоторых других, всегда неотрицательны.

Еще одним способом улучшения качества оценки спектра мощности является *усечение ковариационной функции*, то есть использование при оценивании не всех, а только L первых значений выборочной ковариационной функции. Выбор числа L, однако, связан с неопределенностью, так как при росте L растет и дисперсия оценок спектральной плотности, а при уменьшении L увеличивается их смещение. Обычно рекомендуют использовать $L = K/5$, однако чаще строят выборочную спектральную плотность для трех значений L и выбирают наилучшую в каком-то смысле оценку.

Анализ временных рядов в астрономии чаще всего применяется при исследовании кривых блеска переменных звезд в присутствии ошибок наблюдений. При этом обычно исключают тренд, а периодические составляющие не исключают, так как в присутствии ошибок наблюдений их период или периоды оценить трудно. Но на графике функции спектральной плотности

периодические составляющие сигнализируют о себе появлением максимумов на соответствующих частотах. Часто определение периодов периодических составляющих является целью исследований.

9.4. Сглаживание данных

Выше отмечалось, что временной ряд можно рассматривать как формальный, то есть аргументом *необязательно* должно быть время. Также было отмечено, что временной ряд можно представить как сумму конечного или бесконечного числа гармоник. В некоторых случаях нас интересуют лишь некоторые из большого или даже бесконечного числа гармоник, которые мы рассматриваем как полезный сигнал, а все остальные рассматриваются как мешающий сигнал — шум. Имеются способы подавить сигнал на частотах, на которых явно отсутствует полезная составляющая. В этом случае вклад полезного сигнала в суммарном сигнале повышается. Такой процесс называется **фильтрацией или сглаживанием данных**.

Определение 9.8. *Операторы, используемые для выделения сигналов на определенных интервалах частот и подавления на других, называются* ***числовыми (цифровыми) фильтрами***.

Радикальным способом решения задачи является разложение временного ряда в ряд Фурье (Фурье-преобразование) с последующим урезанием соответствующей функции спектральной плотности. Но есть и более простые способы цифровой фильтрации данных.

Определение 9.9. *В общем случае цифровой фильтр задается выражением* [30]

$$y_k = \sum_{i=-\infty}^{+\infty} c_i\, x_{k-i} + \sum_{i=1}^{+\infty} d_i\, y_{k-i}\,, \tag{9.18}$$

где c_i, d_i — коэффициенты фильтра, которые могут быть функциями входного и выходного сигналов и момента времени, а для линейных и инвариантных во времени фильтров являются константами; y_k — отклик фильтра в k-й точке обработанного фильтром временного ряда; x_k — значение входного сигнала в k-й точке исходного временного ряда.

Если все $d_i = 0$, то цифровой фильтр называется *нерекурсивным*, иначе фильтр называют *рекурсивным*.

Здесь мы рассмотрим простейшую разновидность числовых фильтров — **симметричные нерекурсивные цифровые фильтры**.

Определение 9.10. *Симметричный нерекурсивный цифровой фильтр выражается следующей формулой:*

$$y_k = \sum_{i=-n}^{n} a_{k+i}\, x_{k+i}, \quad a_{k+i} = a_{k-i}\,, \tag{9.19}$$

где k — номер точки во временном ряду; y_k — результат действия фильтра; x_k — значение входного сигнала в k-й точке исходного временного ряда.

Определение 9.11. *Собственное значение оператора цифрового фильтра* (9.18) *при подстановке в него входной функции вида $x_n = \exp(i\,\omega\,n)$ называется* ***передаточной функцией***

фильтра $H(\omega)$. *Умножение передаточной функции на амплитуду входного сигнала на частоте* ω *дает амплитуду выходного сигнала фильтра на этой частоте. Значение передаточной функции показывает, какую долю сигнала на данной частоте* ω *пропускает фильтр, а какую подавляет.*

Вследствие линейности формулы (9.18) по входным данным входной сигнал обязательно присутствует и на выходе фильтра, только помноженный на некоторую функцию, которая является по определению собственным значением входного сигнала.

Для случая монохроматического входного сигнала можно показать, что справедлива следующая теорема [30].

Теорема 9.2 (О наложении частот)**.** *При дискретной равноотстоящей фиксации значений входного сигнала в виде синусоиды любая синусоида произвольной круговой частоты* ω *в точках отсчетов эквивалентна синусоиде с круговой частотой, находящейся в диапазоне значений между 0 и* π, $\nu \in [\,0,\ 1/2\,]$, *где* 2π *— круговая частота, с которой следуют отсчеты сигнала (в системе единиц, где промежуток времени между соседними отсчетами сигнала принят за единицу). Эквивалентность рассматривается в том смысле, что две такие синусоиды имеют одинаковые значения амплитуды в точках отсчета. Таким образом, более высокая, чем* π, *частота будет восприниматься при дискретизации (и исключительно вследствие дискретизации) как более низкая частота из интервала значений от 0 до* π. *Этому эффекту (****наложение*** *—* ***aliasing****) не подвержены только синусоиды с такими частотами, что на каждый период попадает минимум два отсчета (см. рис. 9.2 и теорему Котельникова 9.1).*

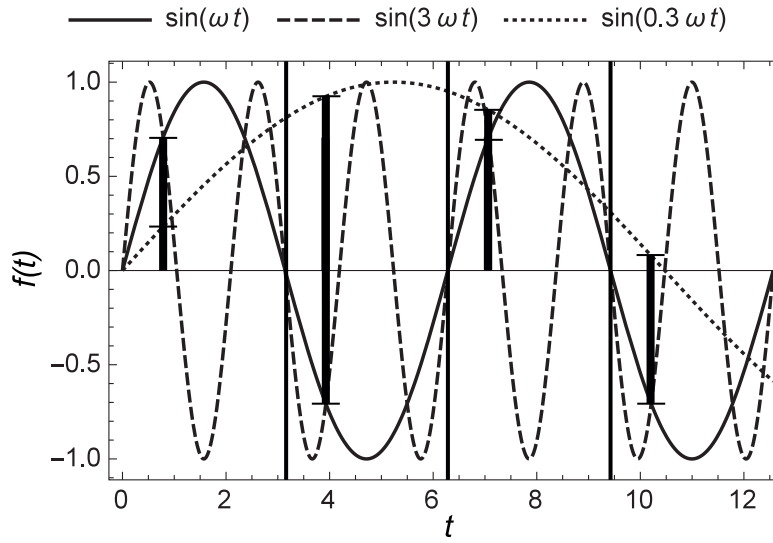

Рис. 9.2. Эффект наложения сигналов при дискретизации. Круговая частота $\omega = 1$. Круговая частота дискретизации $\omega_d = 2\omega = 2$ ($\omega = 2\pi\nu$, $\nu_d = 1/\pi$). Вертикальными непрерывными линиями полной высоты показан период дискретизации входного сигнала $T_d = 2\pi/2 = \pi$ (период, за который берется один отсчет входного сигнала). Вертикальные широкие непрерывные линии показывают реальные моменты взятия отсчетов (следуют через промежуток времени, равный T_d) входных сигналов, в качестве которых выступают синусоидальные сигналы трех различных частот ν ($1/2\pi$, $3/2\pi$, $0.3/2\pi$). Для сигналов с частотами меньшими либо равными половинной частоте дискретизации отсчеты позволяют однозначным образом восстановить исходную синусоиду. Для сигнала с частотой $3/2\pi$ (превышает половинную частоту дискретизации) показан эффект наложения частоты, когда отсчеты сигнала для этой частоты совпадают с отсчетами сигнала для частоты $1/2\pi$ и сигнал с частотой $3/2\pi$ отождествляется с частотой $1/2\pi$

Согласно теореме 9.2 если отсчеты сигнала следуют с круговой частотой 2ω, где ω — некоторая произвольно выбранная круговая частота, то максимальной круговой частотой синусоидального сигнала, для которой *не будет* наблюдаться эффект наложения частот, будет круговая частота $\omega_{\max} = \omega$.

Если перенормировать на единицу промежуток времени между соседними отсчетами сигнала Δt, равный периоду T_d дискретизации сигнала

$$T_d = \Delta t = 1\,, \tag{9.20}$$

разделив все интервалы времени в рассматриваемой системе на T_d, то тогда круговая частота дискретизации и максимальная допустимая частота сигнала равны

$$\omega_d = 2\omega = \frac{2\pi}{T_d} = 2\pi\,, \tag{9.21}$$

$$\omega_{\max} = \omega = \frac{\omega_d}{2} = \pi\,. \tag{9.22}$$

От круговой частоты ω можно перейти к обычной частоте $\nu = \omega/2\pi$. Тогда в системе единиц, где $T_d = \Delta t = 1$,

$$\nu_d = 1\,, \tag{9.23}$$

$$\nu_{\max} = \frac{1}{2}\,, \tag{9.24}$$

что по сути указывает, что в относительной системе единиц максимальная частота входного сигнала не может превышать половины частоты дискретизации для того, чтобы избежать появления эффекта наложения сигналов.

Переход в систему единиц, где $T_d = \Delta t = 1$, позволяет рассматривать свойства цифровых фильтров, в реальности работающих на различных частотах, в одной масштабной шкале. Далее рассмотрение примеров цифровых фильтров ведется в системе единиц, где $T_d = \Delta t = 1$. Поведение фильтров анализируется только для частот входных сигналов $\nu \leqslant 1/2$, поскольку подача на вход цифрового фильтра сигнала с частотами, превышающими $1/2$, приведет при дискретизации такого сигнала к эффекту наложения частот и неизбежному появлению артефактов в выходном сигнале.

Отметим, что самой большой круговой частоте $\omega_{\max}$, численно равной π, *для которой не должен наблюдаться эффект наложения частот*, соответствует расстояние между двумя последовательными точками ряда, а самой низкой частоте соответствует интервал, содержащий столько точек, каков порядок фильтра. Так, для пятиточечного фильтра минимальный интервал — шаг значений аргумента, по которому расположены точки, а минимальная частота соответствует пяти значениям шага. Именно на таких интервалах действует сглаживающая процедура.

Получение коэффициентов a_{k+i} формулы (9.19), которые определяют свойства цифрового фильтра, рассмотрим на простом примере **пятичленного сглаживающего фильтра**, формула для которого имеет вид

$$y_n = a\,x_{n-2} + b\,x_{n-1} + c\,x_n + b\,x_{n+1} + a\,x_{n+2}\,. \qquad (9.25)$$

Рассмотрим воздействие фильтра на монохроматический сигнал с частотой ω, когда мы применяем к нему сглажива-

ющий фильтр. Пусть

$$x_n = \exp(i\,\omega\,n), \tag{9.26}$$

где i в данном случае есть мнимая единица. Исследуем выходной сигнал, получаемый после применения фильтра. Подставим монохроматический сигнал (9.26) в фильтр (9.25) и получим передаточную функцию

$$H(\omega) = 2\,a\cos(2\,\omega) + 2\,b\cos(\omega) + c\,. \tag{9.27}$$

Если бы не было условия симметричности, в последнем выражении присутствовали бы синусные члены с мнимыми коэффициентами.

Теперь займемся подбором коэффициентов a, b и c. Ответ на вопрос, как находить значения этих коэффициентов, зависит от того, какие частоты мы хотим пропустить, а какие подавить. Рассмотрим расчет **фильтра низких частот**, который подавляет высокочастотную составляющую сигнала. Такие фильтры на практике применяются наиболее часто, поскольку быстрые изменения функции обычно отождествляются со случайными ошибками, и именно их стараются подавить. Начнем с наложения двух условий на передаточную функцию: для низкочастотного конца функции потребуем, чтобы $H(0) = 1$, то есть чтобы сигнал на самой низкой частоте не менялся под действием фильтра. Для высокочастотного конца потребуем выполнения условия $H(\pi) = 0$, чтобы высшие частоты фильтром полностью подавлялись. Эти два условия для выражения (9.27)

дают

$$\begin{aligned} H(0) &= 2\,a + 2\,b + c = 1, \\ H(\pi) &= 2\,a - 2\,b + c = 0. \end{aligned} \tag{9.28}$$

Решив эту систему из двух уравнений, получим

$$b = \frac{1}{4}\,; \qquad c = \frac{1}{2} - 2\,a\,, \tag{9.29}$$

таким образом, задача сводится к однопараметрическому семейству фильтров

$$H(\omega) = 2\,a\cos(2\omega) + \frac{1}{2}\cos(\omega) - 2\,a + \frac{1}{2}\,. \tag{9.30}$$

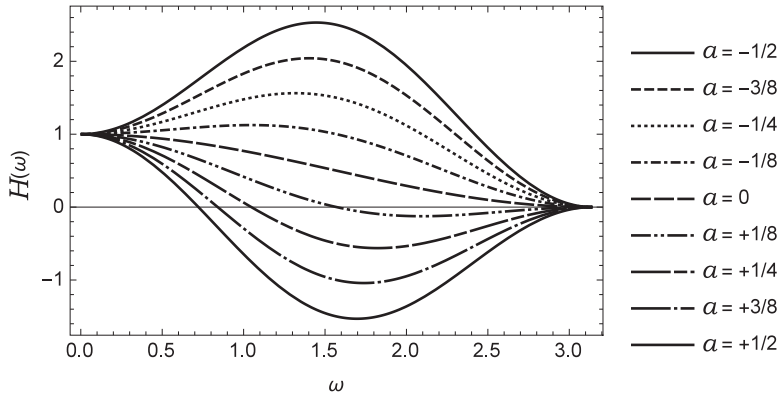

Рис. 9.3. Семейство передаточных функций для однопараметрического пятичленного сглаживающего фильтра (9.25) для диапазона параметра $a \in \left[-\frac{1}{2},\ \frac{1}{2}\right]$

Построив графики функции $H(\omega)$ для разных значений параметра, мы можем выбрать наилучший, по нашему мнению, фильтр (рис. 9.3).

Еще один путь — наложить *добавочное* условие на функцию $H(\omega)$. Например, таким условием может быть более высокий порядок касания передаточной функцией асимптот при крайних значениях частот, то есть либо $H''(0) = 0$, либо $H''(\pi) = 0$, где штрих означает взятие производной (первые производные при этом оказываются равными нулю автоматически, так как взятие производной превращает косинусы в синусы). Добавив одно из этих условий к системе (9.28), мы полностью определим коэффициенты фильтра. Пусть мы используем первое условие. В этом случае получаем значения коэффициентов $a = -1/16$, $b = 1/4$, $c = 5/8$. Полностью фильтр имеет вид

$$y_n = \frac{1}{16}\left(-u_{n-2} + 4\,u_{n-1} + 10\,u_n + 4\,u_{n+1} - u_{n+2}\right) . \qquad (9.31)$$

Теперь, подставляя в эту формулу последовательно значения функции, можно вычислять сглаженные значения. Построение передаточной функции для данного фильтра оставим читателю в качестве упражнения.

Указанным способом можно построить симметричные нерекурсивные цифровые фильтры любого порядка. При этом следует учитывать, что **идеальный фильтр** низких частот, казалось бы, должен иметь *прямоугольную передаточную функцию*, когда все низшие частоты передаются совершенно без искажений, а все частоты выше частоты среза фильтра полностью подавляются. Поэтому при выборе коэффициентов следо-

вало бы добиваться такой передаточной функции для выбранного порядка фильтра, чтобы эта функция максимально близко напоминала функцию-ступеньку. Однако анализ свойств идеального фильтра показывает, что его применение приводит к сильным артефактам в обработанном сигнале (появлению «звона» на фронтах сигнала и эхо-эффектам). В реальности применяют фильтры с более гладкими передаточными характеристиками. За подробностями отсылаем читателя к специальной литературе по цифровой обработке сигналов.

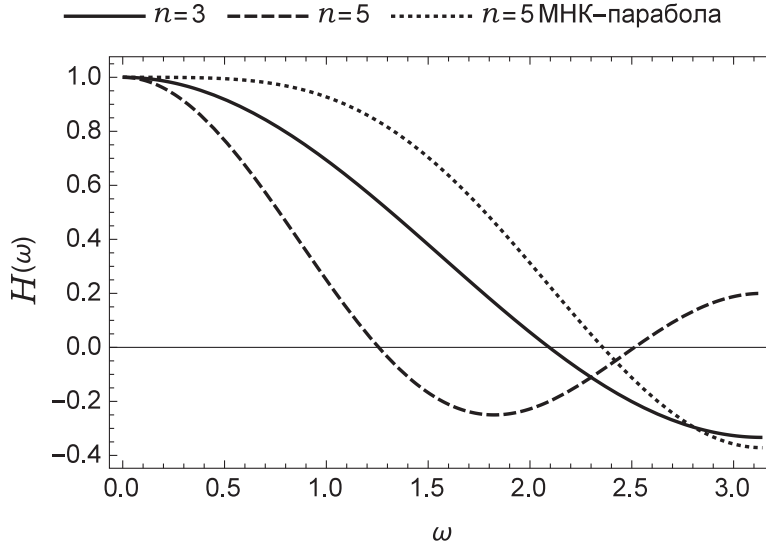

Рис. 9.4. Передаточные функции для простого трехточечного, пятиточечного фильтров и для сглаживания параболами, построенными по пяти точкам методом МНК

На рис. 9.4 показаны передаточные функции фильтров, часто применяемых на практике при обработке данных наблю-

дений в астрономии. Передаточные функции построены для простых трех- и пятиточечных усредняющих фильтров и для усреднения параболой, построенной по пяти точкам исходного ряда методом наименьших квадратов. Формулы для передаточных функций указанных фильтров приведены ниже.

Осреднение по трем точкам:

$$H(\omega) = \frac{1}{3}\left(2\cos(\omega) + 1\right) . \tag{9.32}$$

Осреднение по пяти точкам:

$$H(\omega) = \frac{1}{5}\left(2\cos(2\omega) + 2\cos(\omega) + 1\right) . \tag{9.33}$$

Осреднение по пяти точкам МНК-параболой:

$$H(\omega) = \frac{1}{35}\left(24\cos(\omega) - 6\cos(2\omega) + 17\right) . \tag{9.34}$$

Показанным выше способом можно построить также фильтры **высоких частот** и **полосовые фильтры**. Для построения высокочастотного фильтра по имеющемуся низкочастотному фильтру достаточно использовать разность вида $x_k - y_k$ в качестве высокочастотного фильтра.

Полосовые фильтры используются для выделения из сигнала определенной гармоники или суммы близких по частотам гармоник. Иногда в качестве полосового фильтра используют последовательное применение фильтров низких и высоких частот.

На рис. 9.5 показан пример функции $y = \sin(x)$, зашумленной нормально распределенными ошибками с дисперсией, равной единице. Число точек на графике равно 100. На рис. 9.6

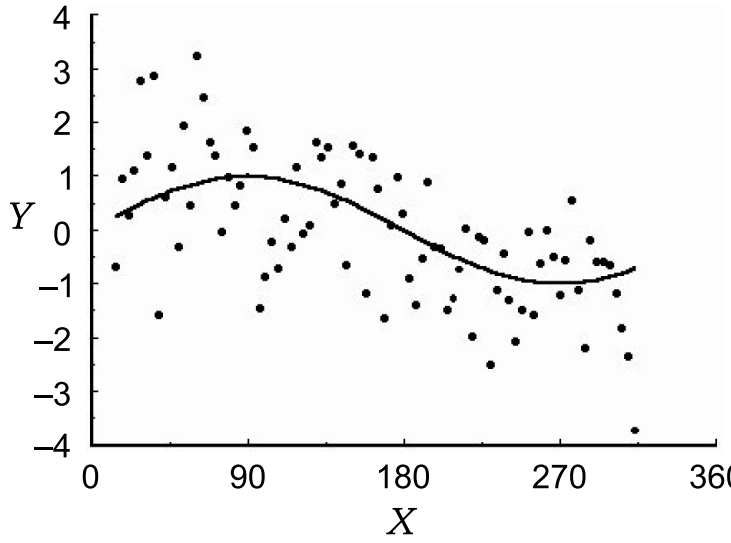

Рис. 9.5. Функция $y = \sin(x)$, зашумленная нормально распределенными ошибками с дисперсией, равной единице

этот же ряд эквидистантных точек сглажен девятиточечным фильтром, задаваемым формулой

$$\begin{aligned} y_i = {} & 0.36\, x_i + \\ & + 0.28125\,(x_{i-1} + x_{i+1}) + 0.093(3)\,(x_{i-2} + x_{i+2}) - \\ & - 0.03125\,(x_{i-3} + x_{i+3}) - 0.023(3)\,(x_{i-4} + x_{i+4})\,. \end{aligned} \qquad (9.35)$$

Так как число точек ряда, показанного на рис. 9.5, равно 100, то минимальный период равен 11 % от всего интервала 360°, на котором задана функция. Несмотря на большую дисперсию ошибок, сглаживающее действие фильтра ясно видно. Большую степень сглаживания можно получить либо ис-

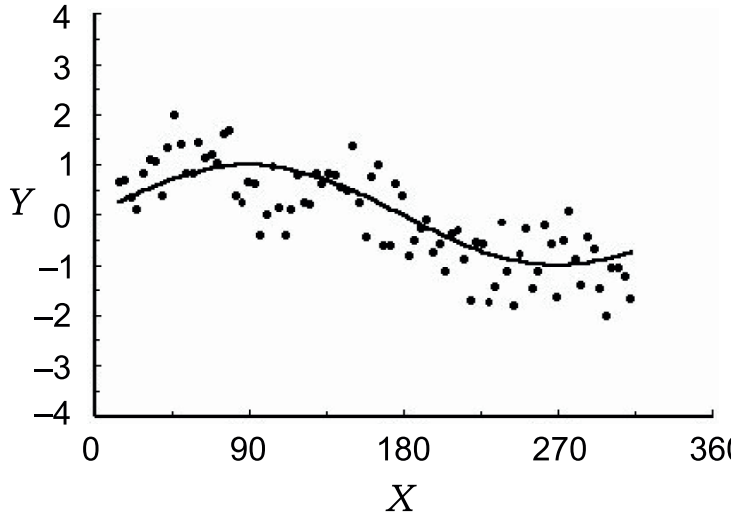

Рис. 9.6. Результат сглаживания девятиточечным фильтром сигнала, показанного на рис. 9.5

пользуя фильтр более высокого порядка, либо, если это неудобно, последовательным применением одного и того же фильтра. Двукратному применению одного фильтра соответствует передаточная функция, равная произведению двух передаточных функций примененных фильтров.

Отметим, что цифровую фильтрацию следует использовать с осторожностью в случае, когда от наблюдаемой зашумленной функции следует ожидать резких скачков. Скачки соответствуют присутствию высоких частот, которые будут подавляться фильтром, и все скачки, в том числе и узкие максимумы или минимумы, будут сглаживаться.

Здесь мы рассмотрели простейший вид цифровых фильтров. Значительно более эффективные фильтры получаются

с помощью преобразования Фурье исходного ряда, усечения результата на нужной частоте и применения обратного преобразования Фурье. Однако эта процедура требует значительно больших усилий, поэтому симметричные цифровые фильтры также находят широкое применение.

9.5. Вейвлет-анализ

Основная идея вейвлет-преобразования отвечает специфике многих временных рядов, демонстрирующих эволюцию во времени своих основных характеристик — среднего значения, дисперсии, а также периодов, амплитуд и фаз гармонических составляющих.

В том случае когда периодическая составляющая случайного процесса ведет себя сложным образом, например, когда периодический сигнал то появляется, то исчезает или меняет свою частоту со временем, рассмотренные выше методы анализа с помощью Фурье-преобразования работают недостаточно уверенно.

Приведем пример. Преобразование Фурье для функции $f(t) = A\cos(\omega_0 t)$ имеет вид

$$F(\omega) = A\sqrt{\frac{\pi}{2}}\left(\delta(\omega - \omega_0) + \delta(\omega + \omega_0)\right) \tag{9.36}$$

для случая, когда прямое преобразование Фурье определено как

$$\hat{\mathrm{F}}\left(f(t)\right) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(t)\exp(i\,\omega\,t)\,dt\,, \tag{9.37}$$

а дельта-функция Дирака определяется стандартным образом:

$$\delta(\omega) = \begin{cases} +\infty, & \text{если } \omega = 0, \\ 0, & \text{если } \omega \neq 0, \end{cases}$$
$$\int_{-\infty}^{+\infty} \delta(\omega)\, d\omega = 1\,. \tag{9.38}$$

Этот пример демонстрирует замечательное свойство преобразования Фурье собирать в точку, «размазанную» по временному ряду, информацию о периодических изменениях функции при переходе из временной области в частотную. Если параметры сигнала меняются с течением времени (зависят от пространственных переменных, как в задачах обработки изображений), то можно рассмотреть преобразование Фурье не на всей временной оси, а на ее частях. Например, такой подход обеспечивает **преобразование Габора**

$$GT(\nu,\, a,\, b) = \int_{-\infty}^{+\infty} f(t) \exp\left(-\frac{(t-b)^2}{a^2}\right) \exp\left(-2\pi\, i\, \nu\, t\right)\, dt\,, \tag{9.39}$$

в котором гауссово окно $\exp\left(-(t-b)^2/a^2\right)$ вырезает по оси времени определенную область. При этом ширина окна, задаваемая параметром a, постоянна. Параметр b задает позицию на оси времени, над которой вывешивается окно. Если *сделать ширину окна переменной*, такой, чтобы каждый времен-

ный компонент анализировался со степенью детальности, которая соответствует его масштабу, то мы приходим к **вейвлет-анализу**, то есть к анализу сигнала с привязкой к определенным промежуткам времени (от англ. wavelet — «всплеск»).

Определение 9.12. ***Интегральным вейвлет-преобразованием*** *функции* $f(t)$ *называется преобразование вида*

$$W(a,\ b) = \frac{1}{|a|^{1/2}} \int\limits_{-\infty}^{+\infty} f(t)\, \psi\left(\frac{t-b}{a}\right) dt\,. \tag{9.40}$$

Величины $W(a,\ \ b)$ *обычно называются* ***вейвлет-коэффициентами****. После преобразования анализируется поведение вейвлет-коэффициентов на плоскости* $(a,\ b)$*, где* b *играет роль частоты в Фурье-анализе. Функция* $\psi(t;\ \ a,\ \ b)$ *называется анализирующим, базисным или* ***материнским вейвлетом****. Параметр* a *определяет размер окна и называется* масштабом*. Существует и преобразование, обратное к* (9.40).

Наиболее часто в настоящее время используются два вида материнских вейвлетов. Первый называется **MHAT-вейвлетом** (от mexican hat — сомбреро). Его математическое выражение получается путем двукратного дифференцирования выражения для плотности нормального распределения, в результате получается

$$\psi(t) = \left(1 - \frac{t^2}{a^2}\right) \exp\left(-\frac{t^2}{2\,a^2}\right), \tag{9.41}$$

где масштаб определяется величиной a.

Второй часто используемый материнский вейвлет — так называемый **вейвлет Морле**. Он записывается в виде

$$\psi(t) = \left(\exp(i\,k\,t) - \exp\left(-\frac{k^2\,a^2}{4}\right)\right)\exp\left(-\frac{t^2}{a^2}\right), \qquad (9.42)$$

где обычно выбирают $a^2 = 2$ и $k = 2\pi$.

Обычно вейвлет Морле используют для анализа временных рядов [31], тогда как МНАТ-вейвлет — для сглаживания данных. Выражение (9.40) показывает, что вейвлет-преобразование можно считать цифровой фильтрацией, при этом передаточная функция фильтра имеет вид

$$H(\omega) = \sqrt{a}\,\psi(a\,\omega)\,. \qquad (9.43)$$

При этом $H(0) = 0$, а максимум лежит на частоте $\omega_{\max} = c/a$, причем $c = \sqrt{2}$ для МНАТ-вейвлета и $c = k = 2\,\pi$ для вейвлета Морле. Для МНАТ-вейвлета выражения (9.41) и (9.43) показывают, что мы имеем дело с полосовым фильтром, ширина полосы которого определяется величиной масштаба a. Это свойство позволяет использовать МНАТ-вейвлет для анализа распределения плотности различных объектов как в проекции на небесную сферу или какую-либо плоскость (например, плоскость Галактики) или даже в пространстве.

В качестве примера использования вейвлет-сглаживания при исследовании распределения молодых звездных объектов в проекции на плоскость Галактики рассмотрим такое распределение для переменных типа δ Цефея.

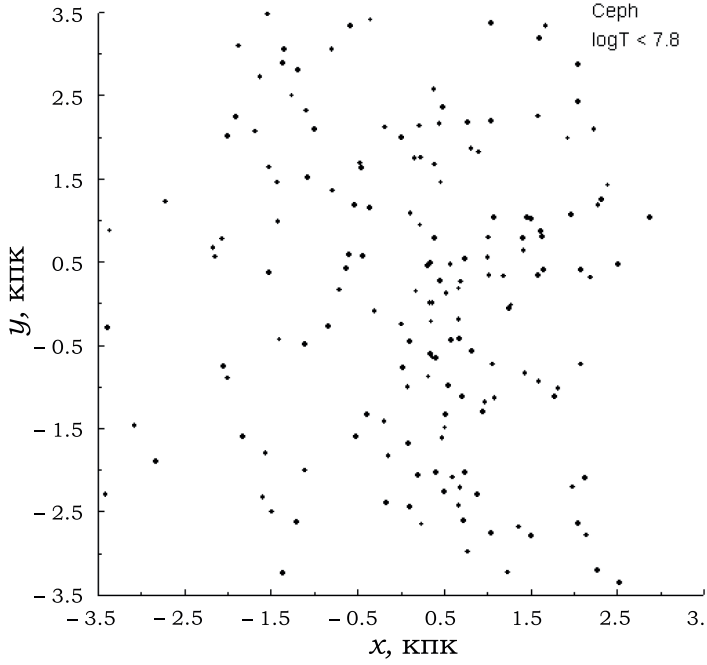

Рис. 9.7. Распределение цефеид в проекции на плоскость Галактики

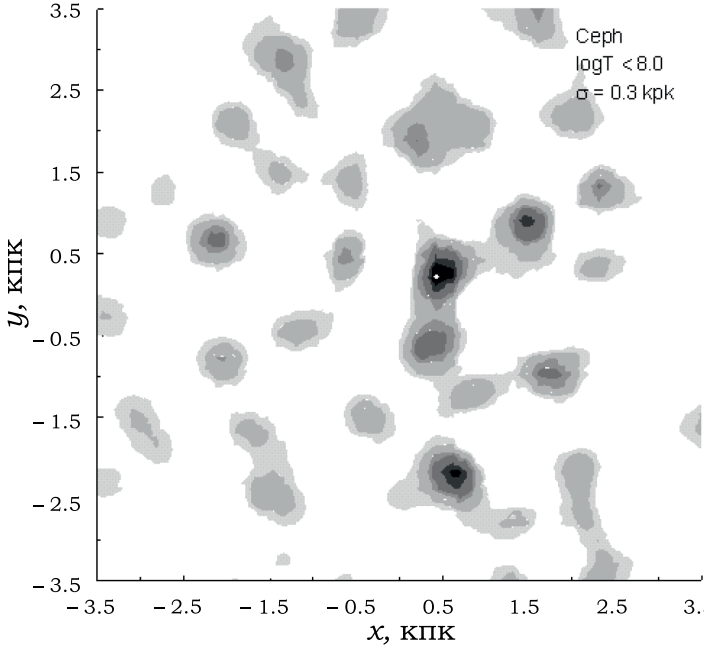

Рис. 9.8. Распределение цефеид в проекции на плоскость Галактики, сглаженное применением формулы (9.44) ($\sigma = 0.3$ кпк)

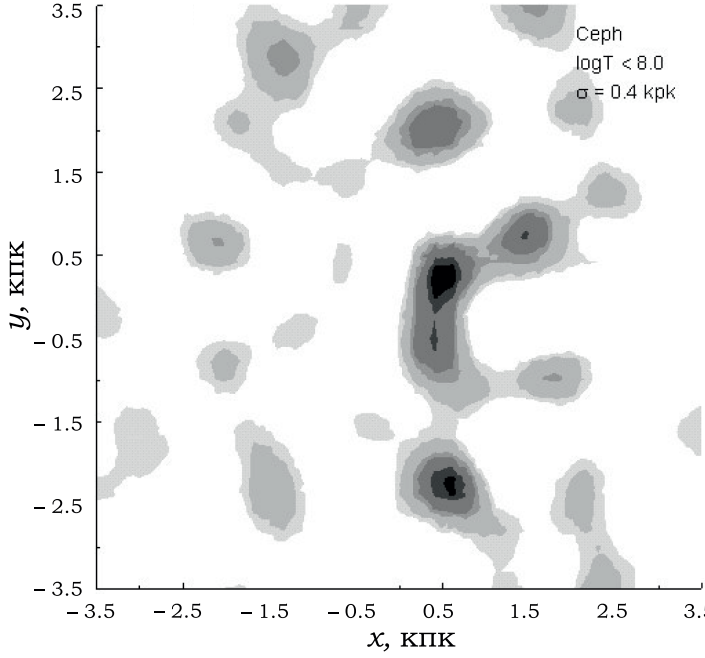

Рис. 9.9. Распределение цефеид в проекции на плоскость Галактики, сглаженное применением формулы (9.44) ($\sigma = 0.4$ кпк)

Для вычисления вейвлет-коэффициентов для дискретного двумерного множества точек использована, с применением MHAT-вейвлета, формула

$$W(k,\ j,\ \sigma) = \sum_{i=1}^{N} \left(2 - \frac{(k-x_i)^2 + (j-y_i)^2}{\sigma^2}\right) \times \\ \times \exp\left(-\frac{(k-x_i)^2 + (j-y_i)^2}{2\,\sigma^2}\right). \tag{9.44}$$

Эта формула получается из (9.40) и (9.41) путем обобщения на двумерный случай и замены интегралов на соответствующие интегральные суммы. В (9.44) k и j являются новыми пространственными переменными; x_i и y_i — координаты цефеид. На рис. 9.7 показано распределение цефеид в проекции на плоскость Галактики, а на рис. 9.8 и 9.9 — аналогичное распределение, но сглаженное применением формулы (9.44), *распределение вейвлет-коэффициентов*. При этом последние два рисунка получены для двух разных значений масштаба. На рисунках центр Галактики — справа, направление галактического вращения — вверх.

Рисунки ясно показывают, что сглаживание позволяет гораздо более удобным образом увидеть структуры в распределении звезд в плоскости Галактики, в частности, выделить элементы спиральной структуры.

Вейвлет-сглаживание приводит к более удобному представлению данных, так как является фильтрацией данных с помощью полосового фильтра. При этом в определенной степени подавляется влияние селекции по расстоянию от Солнца (уменьшение плотности точек вследствие уменьшения вероятности от-

крытия объекта с ростом расстояния от Солнца) и сглаживается влияние отдельных объектов, что соответствует отфильтровыванию самых высоких частот.

10. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ В ЗВЕЗДНОЙ СТАТИСТИКЕ

10.1. Генерирование последовательностей псевдослучайных чисел

Из предыдущих глав мы узнали, что зачастую статистические задачи достаточно сложны, так что влияние ошибок наблюдений на оценивание параметров и влияние отклонений от допущений, при которых применяются те или иные статистические методы, может быть непредсказуемым. Часто затруднительно выбрать наиболее эффективный метод исследования изучаемого явления. Во многих случаях решению задачи может помочь численный эксперимент, предшествующий решению основной задачи. При этом численная модель практически никогда не воспроизводит изучаемое явление точно, но может ограничить круг основных моментов, которые необходимо иметь в виду при выборе метода и решении задачи.

Иногда в литературе численные эксперименты называют **методом Монте-Карло**, смешивая несколько различных методов решения задач вычислительной математики и численные эксперименты. Мы предпочитаем отделять численные

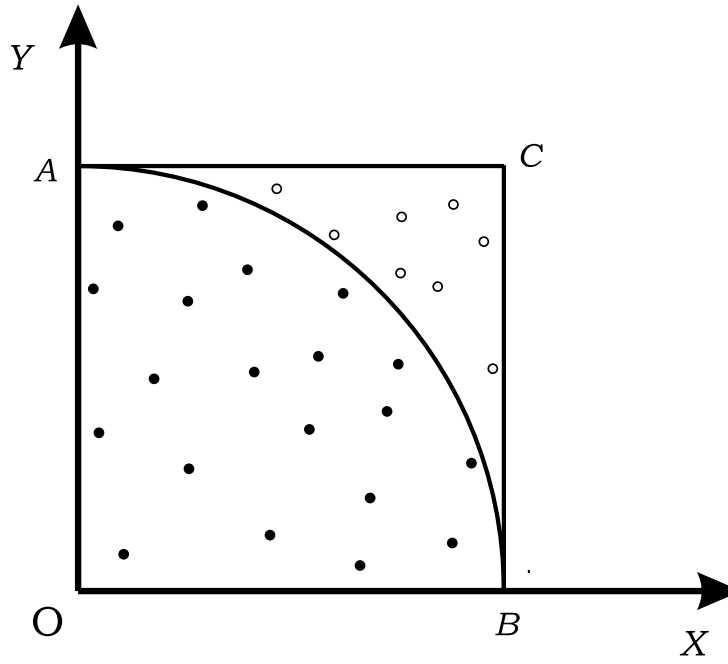

Рис. 10.1. Оценка значения числа π методом Монте-Карло

статистические эксперименты от метода Монте-Карло, оставляя за последним решение задач вычислительной математики, где неслучайные явления исследуются с применением случайных величин.

В качестве простейшего примера применения метода Монте-Карло часто приводят определение величины числа π. Идея решения задачи методом Монте-Карло понятна из рис. 10.1. Площадь сектора круга OAB равна $\pi r^2/4$ или, для единичного радиуса, $\pi/4$. Площадь квадрата $OACB$ равна в этом случае единице. Пусть мы имеем возможность создать длинную после-

довательность случайных чисел, распределенных равномерно на интервале $[0,\ 1]$. Образуем из этого ряда координаты случайных точек $(x,\ y)$. Эти точки всегда будут попадать в квадрат $OABC$ и иногда в сектор OAB. Отношение числа случайных точек, попавших в сектор, то есть таких, что $x^2 + y^2 < 1$, к общему числу использованных случайных точек равно, очевидно, $\pi/4$. Используя случайную последовательность все большей длины, мы можем получить оценку числа π с любой заданной точностью.

Приведенный пример характерен для решения задач методом Монте-Карло. Задача здесь вполне детерминирована, не содержит случайных величин, и только результат решения задачи можно рассматривать как случайную величину: если мы проведем много испытаний с разными последовательностями случайных чисел, то результат можем обрабатывать как случайную величину. Со статистическим моделированием метод Монте-Карло роднит использование последовательностей случайных чисел.

Рассмотрим *получение больших последовательностей случайных чисел.* В последнее десятилетие такие последовательности получают с помощью компьютерных программ. Детальное описание алгоритмов приведено в [32]. Так как компьютерные программы всегда дают детерминированный результат, то такие числа называют **псевдослучайными**. При этом обычно получают равномерно распределенные на интервале $[0,\ 1]$ псевдослучайные числа. Практически всегда программы для получения таких последовательностей основаны на целочисленном переполнении разрядной сетки компьютера, когда большое целое число умножается на достаточно большую целую констан-

ту; происходит переполнение, при котором определенное число старших разрядов отбрасывается, а оставшаяся часть имеет характер, близкий к случайному. Практически в таких датчиках используется формула

$$Y_{i+1} = A\,Y_i \pmod{2^n - 1} + B\,, \tag{10.1}$$

где A и B — целые константы, определяемые из некоторых соображений (см. ниже); n — натуральное число, задающее период последовательности, обычно не меньшее 32; $\pmod{2^n-1}$ — операция приведения по данному модулю операнда, стоящего слева. Иногда используются и другие подобные формулы. Главное преимущество таких формул — быстрота выполнения операций, так как обычно как в методе Монте-Карло, так и в численных статистических экспериментах используются многие тысячи псевдослучайных чисел.

Доказывается теорема, говорящая о том, что последовательность таким способом определяемых «случайных» чисел обязательно имеет период, то есть через некоторое количество чисел новые числа начинают повторять предыдущие. Это является следствием детерминированности всех компьютерных операций. При этом имеется оценка только на максимальное значение периода: это 2^P, где P — разрядность компьютера. Константы в формуле (10.1) определяются так, чтобы длина последовательности для любого начального целого числа была как можно больше. Подробно о выборе констант можно прочитать в книге Дж. Форсайта [33], там же можно найти и универсальную программу — датчик псевдослучайных чисел. Так как качество последовательности псевдослучайных чи-

сел может быть разным, после написания программы — датчика случайных чисел результаты ее работы обязательно должны проверяться. Основными способами проверки вычисляемого ряда псевдослучайных чисел являются проверка распределения этих чисел на равномерность и построение автокорреляционной функции. Последняя показывает, не коррелированы ли между собой разные группы псевдослучайных чисел, и позволяет увидеть периодичность их появления. При проверке обязательно используются большие последовательности длиной в несколько миллионов чисел.

Из равномерно распределенных псевдослучайных чисел обычно нетрудно получить числа, распределенные по иным законам. О преобразовании распределений говорилось в главе 1. В тех случаях когда трудно получить обратную функцию для функции распределения, используются специальные численные методы. Примером такого случая является нормальное распределение. Приведем два способа получения последовательностей псевдослучайных чисел, распределения которых близки к нормальному распределению с нулевым средним и единичной дисперсией $N(0,\ 1)$.

1. Метод Мюллера. В этом методе два последовательных *равномерно распределенных* числа R_1 и R_2 преобразуются в два *нормально распределенных* N_1 и N_2 по формулам

$$\begin{aligned} N_1 &= \sqrt{-2\ln R_1}\,\cos(2\,\pi R_2)\,, \\ N_2 &= \sqrt{-2\ln R_1}\,\sin(2\,\pi R_2)\,, \end{aligned} \tag{10.2}$$

   причем получающиеся пары чисел некоррелированы.

2. Следующий метод работает на многих компьютерах несколько быстрее, чем предыдущий. В этом методе из двух равномерно распределенных случайных чисел R_1 и R_2 сначала образуются числа $V_1 = 2R_1 - 1$ и $V_2 = 2R_2 - 1$, и, если $S = V_1^2 + V_2^2 > 1$, такие пары отбрасываются, а для прошедших отбор пар вычисляется *нормально распределенное* число

$$N = V_1 \sqrt{-\ln S/S}\,. \tag{10.3}$$

10.2. Численный эксперимент в звездной кинематике

Рассмотрим практическое применение численного эксперимента в звездной астрономии на примере определения кривой частоты вращения диска Галактики по лучевым скоростям ОВ-звезд. Конкретизация объектов исследования необходима потому, что для моделирования ситуации требуется иметь оценки ошибок определения расстояний до объектов и лучевых скоростей. Понятно, что точность определения расстояний до рассеянных звездных скоплений или цефеид гораздо выше, чем до отдельных ОВ-звезд. То же самое можно сказать и о лучевых скоростях, так как лучевые скорости скоплений определяются как средние лучевые скорости группы звезд. Рассмотрим *формулу Боттлингера*, связывающую кривую вращения диска Галактики с наблюдаемой лучевой скоростью, исправленной за движение Солнца в пространстве:

$$v_R = R_0(\omega_0 - \omega)\sin l \cos b\,, \tag{10.4}$$

где R_0 — расстояние Солнца от оси вращения Галактики; ω_0 — частота вращения (угловая скорость) диска Галактики на расстоянии Солнца от оси вращения; $\omega = \omega(R)$ — частота вращения диска Галактики на расстоянии R от ее оси вращения, то есть кривая угловой скорости вращения диска Галактики. Получение этой функции является одной из главных задач звездной кинематики. Из выражения (10.4) можно записать

$$R_0 \left(\omega - \omega_0\right) = \frac{v_R}{\sin l \cos b}\,, \tag{10.5}$$

эту функцию обычно называют *функцией Камма—Паренаго* или просто *функцией Камма*. Она отличается от кривой линейной скорости вращения диска Галактики на величину постоянной поправки — линейной скорости вращения Галактики в точке Солнца ω_0. Для принятой величины R_0 значения этой функции строят в зависимости от расстояния звезды, имеющей лучевую скорость v_R, от оси вращения Галактики $R = \sqrt{R_0^2 + r^2 \cos^2 b - 2R_0\, r \cos l \cos b}$, где r — расстояние от звезды до Солнца. Во многих практических задачах галактическая широта может быть принята равной нулю.

Чтобы смоделировать влияние случайных ошибок в данных наблюдений, а в этой задаче используются определяемые из наблюдений *расстояния до звезд* и *лучевые скорости*, необходимо построить модельную выборку, относительно которой и будет рассматриваться влияние ошибок. Для этого следует использовать известную кривую вращения, которую приближенно можно задать из простой модели Галактики например. В частности

можно взять кривую вращения в виде

$$\omega(R) = \frac{M_1}{1 + a_1 R^2} + \frac{M_2}{1 + a_2 R^2} \tag{10.6}$$

с параметрами, подобранными для приближения известной кривой вращения. Нам нет необходимости работать с определенными числами, так как мы рассматриваем лишь принципы статистического моделирования. По той же причине будем рассматривать моделирование только качественно.

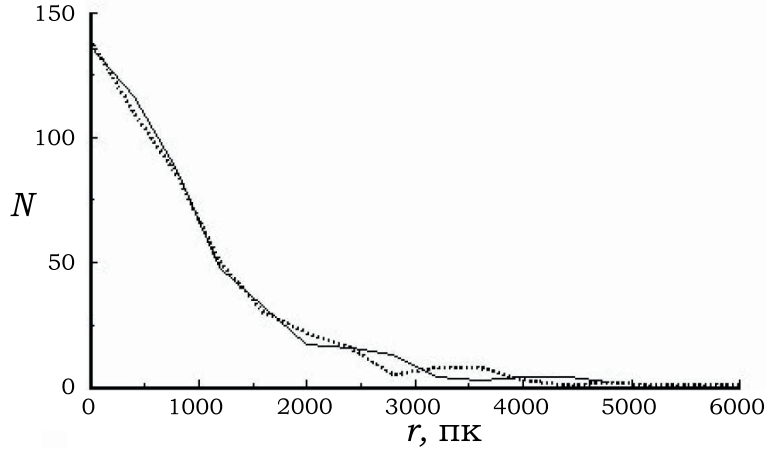

Рис. 10.2. Модельное экспоненциальное распределение выборки звезд по расстоянию от Солнца (непрерывная линия) и распределение с введенными ошибками в модули расстояний (точечная линия). Расстояние r измеряется в парсеках

Начнем моделирование с создания искусственной выборки. Для этого зададим достаточно большое число точек, служащих звездами выборки, для которых с помощью датчика случайных чисел зададим расстояния от Солнца r и галакти-

ческие долготы l. При этом галактические долготы должны быть распределены *равномерно* в интервале $[\,0,\ 360^\circ\,]$, для чего создадим массив равномерно распределенных случайных чисел, распределенных на единичном интервале, и умножим все числа массива на 360. Расстояния от Солнца будем задавать так, чтобы они были распределены с экспоненциальной плотностью $\sim \exp(-d\,r)$, которая обычно хорошо приближает реальные выборки, при этом постоянная d подбирается для согласия с имеющейся выборкой реальных объектов.

Далее следует перейти к интегральной функции распределения, которая равна $F = 1 - \exp(-d\,r)/d$, а затем выразить r через F: $r = -\ln\left(d\,(1-F)\right)/d$. Если теперь подставлять вместо F в это выражение равномерно распределенные псевдослучайные числа, мы получим числа, распределенные экспоненциально. Умножив эти числа на постоянную, которую выберем вместе с d так, чтобы распределение было похоже на распределение имеющейся реальной выборки, получим распределение расстояний до модельных точек, напоминающее распределения обычно используемых в кинематических исследованиях распределения расстояний O и B звезд. Оно показано на рис. 10.2 сплошной линией.

Теперь введем ошибки, причем ошибки следует вводить в модули расстояния, а не в сами расстояния, так как с ошибками определяются из наблюдений звездные величины, а не расстояния. Поэтому для всех расстояний модельных звезд нашей выборки вычислим модули расстояния и прибавим к ним нормально распределенные случайные числа с дисперсией 0.6^m, которая характерна для ошибок оценок абсолютных звездных величин O и B звезд, а затем вновь перейдем к расстояниям.

Частотное распределение полученных таким образом расстояний показано на рис. 10.2 пунктиром. Как видим, распределение после введения ошибок почти не изменилось, среднее значение расстояния увеличилось при этом всего на 4 %. Единственное различие — продление правого хвоста распределения: если у истинных расстояний распределение простиралось приблизительно до 5 000 пк, то у распределения с ошибками появились звезды с расстояниями до 6 000 пк. Детали методики исправления наблюдаемых распределений подробно описаны в [34].

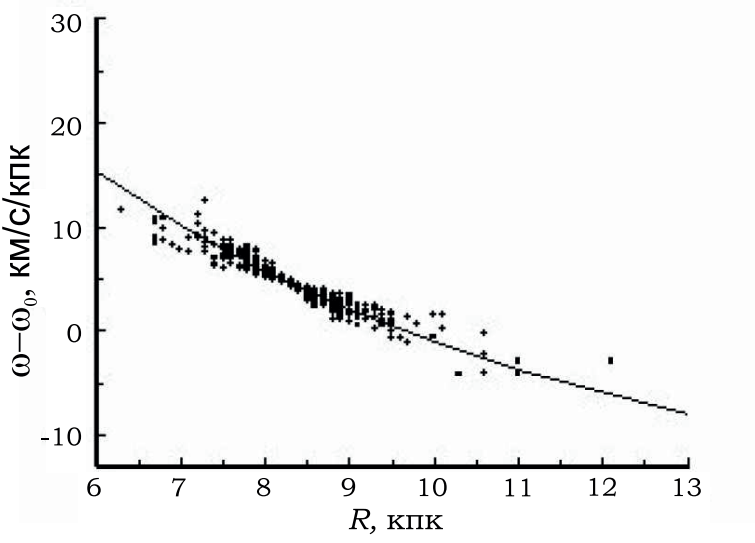

Рис. 10.3. Разность угловых скоростей на галактоцентрических расстояниях R и R_0 для модельных звезд выборки, отягощенная ошибками модулей расстояния (точки); та же разность угловых скоростей для принятой модели вращения Галактики (непрерывная линия). Галактоцентрические расстояния даны в килопарсеках

Следующим шагом эксперимента будет оценивание влияния ошибок на значения функции Камма. Для этого мы должны снабдить каждую звезду выборки лучевой скоростью, которую мы вычисляем, сначала определяя для каждой звезды ее галактоцентрическое расстояние R, затем, вычисляя для данной звезды величину $\omega(R)$ по формуле (10.6); приняв значение $\omega_0 = 25$ км/с/кпк, вычислим лучевую скорость по формуле (10.4). После этого вычисляем галактоцентрические расстояния звезд уже из расстояний от Солнца, отягощенных ошибками, а из формулы (10.5) получаем значения функции Камма. Эта функция представлена на рис. 10.3, где точками показаны значения функции Камма для модельных звезд, а сплошной линией — кривая (10.6), по которой вычислялись лучевые скорости звезд выборки.

Разброс точек на графике вызван исключительно введенными нами ошибками в модулях расстояния искусственных звезд. При этом видно, что точки на краях графика распределены несимметрично относительно гладкой исходной кривой: на правом конце распределения точки лежат преимущественно выше исходной кривой, а на левом конце — ниже, так что, если не учитывать ошибки в расстояниях до звезд, то получаемая кривая вращения будет иметь меньший наклон, чем реальная.

Из слабой зависимости распределения расстояний от ошибок наблюдений и появившейся некоторой асимметрии оценок функции Камма можно вывести адекватную процедуру исправления влияния ошибок в расстояниях: следует удалить из выборки некоторое количество самых далеких звезд, но само

это количество зависит от формы распределения расстояний звезд реальной выборки. При этом следует запомнить, что иногда случайные ошибки приводят к систематическим ошибкам в результатах.

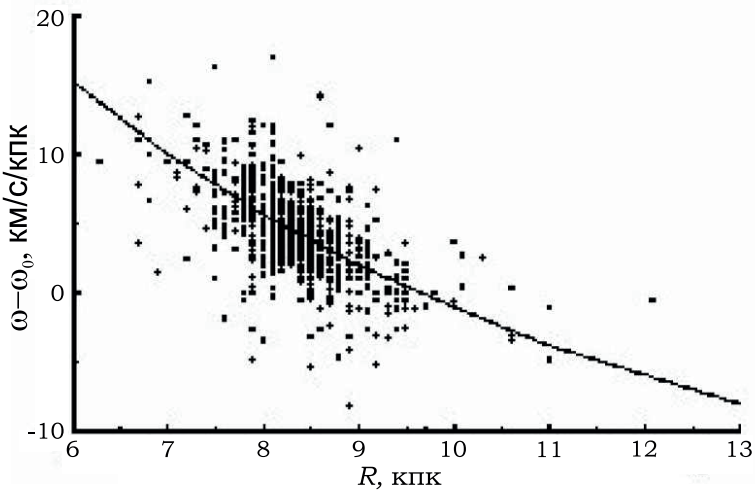

Рис. 10.4. Разность угловых скоростей на галактоцентрических расстояниях R и R_0 для модельных звезд выборки, отягощенная ошибками модулей расстояния, ошибками измерения лучевых скоростей звезд и дисперсией скоростей звезд (точки); та же разность угловых скоростей для принятой модели вращения Галактики (непрерывная линия). Галактоцентрические расстояния даны в килопарсеках

Наконец, введем в лучевые скорости звезд нашей искусственной выборки дисперсию скоростей, присущую молодым звездам диска Галактики (около 10 км/с) с помощью датчика нормально распределенных случайных чисел, а также ошибки измерения лучевых скоростей со среднеквадратичным отклонением, которое можно взять равным средней ошибке из-

мерения по всем звездам реальной выборки (здесь взята величина 3 км/с). После этого вновь вычислим значения функции Камма. Эти значения показаны на рис. 10.4. При этом мы удалили из выборки звезды, имеющие галактические долготы в интервалах $[-8,\ +8^\circ]$ и $[172,\ 188^\circ]$, так как вклад галактического вращения в лучевые скорости таких звезд значительно уступает вкладу пекулярных движений и ошибок измерения лучевых скоростей. На этом эксперимент можно считать законченным.

Проведенный эксперимент дал нам информацию о влиянии ошибок в расстояниях до звезд на оценки функции Камма и подсказал способ их устранения. Учет всех наблюдательных ошибок позволит определить достижимую для данной выборки точность результата и оценить объем и другие характеристики выборки, необходимой для решения поставленной задачи с нужной точностью.

11. СТАТИСТИЧЕСКИЕ МЕТОДЫ В ЗАДАЧАХ ЗВЕЗДНОЙ АСТРОНОМИИ

11.1. Учет эффектов наблюдательной селекции

Одной из основных проблем звездной статистики является получение случайной и *представительной* (репрезентативной) выборки, поскольку почти все эмпирические распределения в звездной астрономии оказываются в большей или меньшей степени искаженными *эффектами наблюдательной селекции*. Универсального метода борьбы с селекцией не существует, поэтому в каждом конкретном случае приходится использовать специальные методы и приемы.

Часто причиной селекции является резкое различие светимостей астрономических объектов, вследствие чего в каталогах, содержащих, как правило, объекты ярче некоторой предельной видимой звездной величины $m_{\lim}$, относительное число абсолютно ярких объектов оказывается сильно завышенным, поскольку чем ярче объект, тем больше расстояние, на котором

его видимая звездная величина станет равной $m_{\lim}$, и тем больше объем пространства, из которого в данный каталог попадают рассматриваемые объекты.

В случае когда известны абсолютные звездные величины изучаемых объектов, учет селекции оказывается наиболее простым и сводится к введению весов, пропорциональных объемам пространства, в которых объекты, ярче данной абсолютной величины M, представлены полностью. Радиус такого сферического объема определяется формулой

$$\lg(r_M) = 0.2\,(m_{\lim} - M) + 1, \tag{11.1}$$

а сам этот способ в звездной астрономии называется **оценкой параметров (или распределений) по сферам полной исчерпанности**.

В общем случае абсолютные звездные величины объектов неизвестны, поэтому при учете селекции приходится жертвовать репрезентативностью выборки, то есть искусственно сокращать выборку так, чтобы ее оставшаяся часть была не искажена эффектами селекции.

Простейшим приближенным способом является *ограничение выборки объектами ярче некоторой видимой звездной величины* m_0, для которых данные каталога *можно считать полными*. Звездную величину m_0 можно найти из графика *функции блеска* $N(m)$, дающей число объектов каталога, имеющих видимую звездную величину ярче m. Если принять, что звезды в окрестностях Солнца распределены равномерно, то функция $N(m)$ должна возрастать экспоненциально в соответствии с *теоремой Зеелигера*.

Теорема 11.1 (Теорема Зеелигера (Hugo von Seeliger)). *Отношение числа звезд некоторой звездной величины $m+1$ к числу звезд звездной величины m в случае однородного распределения в пространстве звезд любых абсолютных звездных величин и отсутствия поглощения света составляет*

$$\frac{N(m+1)}{N(m)} \cong 3.98\,. \tag{11.2}$$

Выполнимость теоремы Зеелигера обычно наблюдается на начальном участке графика $N(m)$, однако, начиная с некоторого значения видимой звездной величины, эмпирическая зависимость $N(m)$ отклоняется от теоретической, что следует связать с эффектом селекции, и в выборку включают только объекты с видимой звездной величиной ярче m_0. Естественно, здесь обязательно следует учитывать особенности пространственного распределения объектов. Например, молодые звезды диска галактики распределены в достаточно тонком слое вокруг плоскости Галактики, и, если рассматриваются сферические объемы с радиусом более 50—70 пк, уменьшение пространственной плотности с ростом z-координаты должно учитываться.

Другим, вероятно более строгим, методом учета селекции является *способ Т. А. Агекяна* или *метод редукции к нулевому расстоянию*. В этом способе не делается никаких предположений о характере распределения объектов, важно только, чтобы изучаемая характеристика не менялась с расстоянием от Солнца.

Пусть нужно оценить некоторый статистический параметр Θ (среднее значение некоторой физической характеристики звезд, число попаданий в некоторый интервал гистограммы и т. д.). Найдем несколько оценок этого параметра $\breve{\Theta}_1$, $\breve{\Theta}_2$, $\ldots$, $\breve{\Theta}_n$, причем для получения каждой k-й оценки используем объекты, находящиеся от Солнца не далее расстояния r_k. Построим теперь график зависимости $\breve{\Theta}_k$ от r_k и проведем через точки полученной зависимости плавную кривую. Проэкстраполировав эту кривую до пересечения с осью ординат, получим в точке пересечения *не искаженное селекцией* значение оценки искомого параметра.

Вместо расстояния можно использовать и звездные величины m_k, однако при этом кривую приходится экстраполировать в минус бесконечность, и поэтому здесь лучше использовать величину 10^{m_k}, которая при $m_k \to -\infty$ стремится к нулю. Можно использовать регрессионный анализ, выбирая для проведения кривой $\breve{\Theta}_k(r_r)$ какую-нибудь простую кривую, например, $\Theta = c/r$. В этом случае можно оценить погрешность экстраполированного значения. Естественно, что ошибка будет достаточно большой, что вызывается как экстраполяцией, так и уменьшением объектов выборок при получении точек с минимальным расстоянием от Солнца.

11.2. Пространственная структура звездного скопления

В этом разделе обсуждаются методы восстановления пространственной структуры звездных скоплений, которые часто применяются в практике звездной статистики.

Видимая звездная плотность в скоплении представляет собой проекцию пространственной плотности на плоскость, касательную к небесной сфере в точке, совпадающей с центром скопления. Задача восстановления пространственной плотности по полученной из наблюдений зависимости звездной плотности от углового расстояния до центра скопления в общем случае, без определенных дополнительных предположений, решения не имеет, что относится и к любой задаче о восстановлении распределения по его маргинальному распределению. Информация о распределении звездной плотности вдоль луча зрения оказывается безвозвратно потерянной. Но решить задачу можно путем использования каких-либо предположений. В частности, можно предположить, что плотность в скоплении распределена сферически симметрично, то есть является функцией только расстояния от центра скопления, что приводит к двум известным методам восстановления пространственной плотности по ее проекции вдоль одной из осей.

Метод Цейпеля

Рассмотрим сначала известный *метод Цейпеля*. Пусть задана прямоугольная система координат $(x,\ y,\ z)$, при этом ось z направлена к наблюдателю. Начало отсчета системы координат поместим, например, в фотометрический центр скопления. Введем обозначения

$$r = \sqrt{x^2 + y^2 + z^2}\,, \tag{11.3}$$

$$\rho = \sqrt{x^2 + y^2}\,. \tag{11.4}$$

Пусть $f(r)$ есть **пространственная плотность** звезд в скоплении, а $F(\rho)$ — **видимая плотность** звезд в плоскости $(x,\ y)$. Функция $F(\rho)$ определяется из звездных подсчетов в кольцевых зонах. В соответствии с формулой (1.9), связывающей многомерное распределение с его частным (маргинальным) распределением, имеем

$$F(\rho) = \int\limits_{-\infty}^{+\infty} f(r)\, dz, \tag{11.5}$$

или, используя соотношение $d\left(r^2\right) = 2z\, dz = 2\sqrt{r^2-\rho^2}\, dz$ (при $\rho = \text{const}$) и учитывая, что r^2 меняется от ρ^2 до $+\infty$, получаем

$$F(\rho) = \int\limits_{\rho^2}^{\infty} \frac{f(r)}{\sqrt{r^2-\rho^2}}\, d\left(r^2\right). \tag{11.6}$$

Это и есть искомое интегральное уравнение, связывающее неизвестную функцию $f(r)$ с получаемой из наблюдений функцией $F(\rho)$. Методики численного решения подобных уравнений хорошо разработаны. Отметим только, что *эта задача относится к классу некорректных*, так как она неустойчива в том смысле, что небольшие ошибки в наблюдаемой функции $F(\rho)$ могут привести к существенным ошибкам в результате, и для ее решения сделаны определенные предположения о симметрии задачи. Для уменьшения влияния ошибок в наблюдаемой функции $F(\rho)$ на результат перед решением *следует сгладить* полученную из подсчетов гистограмму.

Метод Пламмера

Существует и более простой метод — *метод Пламмера*, использующий проекцию пространственного распределения $f(r)$ на ось x, то есть частную одномерную плотность

$$F_1(x) = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(r)\,dy\,dz\,. \tag{11.7}$$

Переходя в плоскости $x = \text{const}$ к полярным координатам $(\varphi,\ \rho)$ и учитывая, что $\rho = \sqrt{y^2+z^2}$, то есть $r^2 = \rho^2 + x^2$, откуда $r\,dr = \rho\,d\rho$, получим интегральное уравнение

$$F_1(x) = 2\pi \int\limits_{x}^{\infty} f(r)\,r\,dr\,, \tag{11.8}$$

где $x > 0$, а верхний предел интегрирования можно приравнять радиусу скопления R, так как за радиусом звездная плотность равна нулю.

Для получения функции $F_1(x)$ необходимо провести звездные подсчеты в узких полосах одинаковой ширины Δx и длины $\Delta y > 2R$. Решение уравнения Пламмера (11.8) достаточно просто. Дифференцируя обе части уравнения по x и разрешая полученное выражение относительно $f(x)$, получаем

$$f(x) = -\frac{1}{2\pi x}\,\frac{dF_1(x)}{dx}. \tag{11.9}$$

Помня о том, что операция численного дифференцирования сильно увеличивает шумы, присутствующие в дифференцируемой функции, результаты звездных подсчетов $F_1(x)$ должны быть тщательно сглажены.

Полученный результат можно использовать и для решения уравнения Цейпеля. Преобразуем выражение (11.7) с учетом (11.5) следующим образом:

$$\begin{aligned} F_1(x) &= \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} f(r)\,dy\,dz = \\ &= \int\limits_{-\infty}^{+\infty}\left(\int\limits_{-\infty}^{+\infty} f(r)dz\right)\,dy = \int\limits_{-\infty}^{+\infty} F(\rho)\,dy\,, \end{aligned} \tag{11.10}$$

где в данном случае $\rho = \sqrt{x^2+y^2}$, то есть если $x = \text{const}$, то $d(\rho^2) = 2y\,dy = 2\sqrt{\rho^2-x^2}\,dy$. Сделав соответствующую подстановку, получим интегральное уравнение

$$F_1(x) = \int\limits_{x^2}^{\infty} \frac{F(\rho)}{\sqrt{\rho^2-x^2}}\,d\left(\rho^2\right)\,, \tag{11.11}$$

связывающее **strip-распределение** (результат подсчетов в полосах) звездной плотности с результатами подсчетов в кольцевых зонах. Теперь достаточно подставить правую часть полученного соотношения в формулу (11.9), чтобы получить решение уравнения Цейпеля:

$$f(x) = -\frac{1}{2\pi x}\,\frac{d}{dx}\int\limits_{x^2}^{\infty} \frac{F(\rho)}{\sqrt{\rho^2-x^2}}\,d\left(\rho^2\right)\,, \tag{11.12}$$

где для реальной задачи верхний предел берется равным R^2. Отметим, что результаты звездных подсчетов *не требуют учета плотности звездного фона*, так как ее постоянное значение не влияет на результат численного дифференцирования.

Метод Пламмера имеет *важное преимущество*: мы можем провести подсчеты, то есть получить стрип-распределение как по оси x, так и по оси y. Сравнивая полученные результаты, можно сделать некоторые выводы о правильности допущения о сферической симметричности изучаемого скопления, а также путем усреднения двух оценок пространственной плотности несколько уменьшить ошибки и грубо оценить погрешность результата.

Для шаровых скоплений, в которых на изображениях центральные части не разрешаются на звезды, используют сканирование изображения узкой щелью, получая одномерное распределение поверхностной яркости, которое с помощью отношения масса—светимость может быть преобразовано в одномерное распределение плотности массы.

11.3. Распределение астрономических объектов и явлений по возрастам и временам жизни

При решении проблем эволюции звездных систем или одиночных звезд, а также при исследовании других изменяющихся со временем объектов и явлений возникает задача оценивания времени жизни (существования) этих объектов или явлений. Непосредственно оценить время жизни астрономических

объектов обычно невозможно, так как оно, как правило, больше времени существования наблюдательной астрономии. В некоторых случаях время жизни объекта может быть вычислено на основе его наблюдаемых характеристик с привлечением теории эволюции, но чаще всего приходится ограничиваться только оценкой распределения объектов по возрастам или по значению некоторого физического параметра объекта (масса, начальная светимость, начальное число звезд в скоплении), однозначно связанного со временем жизни.

Рассмотрим объекты некоторого класса. Пусть T — время жизни конкретного объекта; τ — возраст этого объекта в момент наблюдения; $n(t)$ — скорость рождения, то есть число всех объектов класса, рождающихся за единицу времени; время, отсчитываемое от момента наблюдений в будущее, считается положительным, а в прошлое — отрицательным. Пусть $f_0(T)$ есть начальная плотность распределения объектов рассматриваемого класса по временам жизни; $f(T)$ — наблюдаемое распределение по временам жизни и $g(\tau)$ — наблюдаемое распределение объектов по возрастам.

Найдем сначала связь наблюдаемого и начального распределений по временам жизни T. Рассмотрим объекты с временами жизни от T до $T + dT$. За период времени от t до $t + dt$ рождается

$$dN = n(t)\, dt\, f_0(T)\, dT \tag{11.13}$$

таких объектов. В настоящее время из всех родившихся за время существования Галактики таких объектов наблюдаются, очевидно, только те, что родились менее T лет назад, поэтому

$$f(T)\,dT = f_0(T)\,dT \int\limits_{-T}^{0} n(t)\,dt \tag{11.14}$$

и

$$f(T) = f_0(T) \int\limits_{-T}^{0} n(t)\,dt\,. \tag{11.15}$$

Интегральное уравнение (11.15) связывает наблюдаемую функцию $f(T)$ с двумя неизвестными функциями: $f_0(T)$ и $n(t)$. Чтобы иметь возможность определить все функции, необходимо либо еще одно уравнение, либо нужно принять некоторые предположения об одной из неизвестных функций, например, о скорости рождения объектов рассматриваемого класса. Для звезд галактического поля часто принимают постоянной скорость рождения, то есть $n(t) = n_0 = \mathrm{const}$. Тогда соотношение (11.15) переходит в

$$f(T) = n_0\,T f_0(T)\,, \tag{11.16}$$

откуда следует, что начальная плотность распределения $f_0(T)$ есть

$$f_0(T) = \frac{f(T)}{n_0 T}\,. \tag{11.17}$$

Величину n_0 можно выразить через общее число наблюдаемых объектов данного класса N и их среднее время жизни. Проинтегрируем соотношение (11.16) по времени жизни T:

$$N = \int\limits_0^\infty f(T)\,dT = n_0 \int\limits_0^\infty T f_0(T)\,dT = n_0\,\hat{\mathrm{E}}(T)\,, \tag{11.18}$$

где $\hat{\mathrm{E}}(T) = \int_0^\infty T\, f_0(T)\, dT$ — математическое ожидание времени жизни объекта.

В практических приложениях математическое ожидание времени жизни $\hat{\mathrm{E}}(T)$ заменяется на среднее по выборке, и тогда $n_0 = N/\langle T\rangle$, что дает

$$f_0(T) = \frac{1}{N}\,\frac{\langle T\rangle}{T}\, f(T)\,. \tag{11.19}$$

Функцию $f(T)$, как уже отмечалось, можно оценить на основе наблюдаемого распределения некоторой характеристики объектов данного класса, однозначно связанной с временем жизни. Если величины n_0 и $\langle T\rangle$ неизвестны, то сначала из формулы (11.17) можно найти ненормированную функцию $n_0\, f_0(T) = f(T)/T$, после чего из условия нормировки функции $f_0(T)$ можно найти

$$n_0 = \int\limits_0^\infty \frac{f(T)}{T}\, dT\,, \tag{11.20}$$

а также $\langle T\rangle = N/n_0$.

Рассмотрим теперь связь начального распределения объектов некоторого класса по временам жизни и их наблюдаемого распределения по возрастам. Число наблюдаемых объектов с возрастами от τ до $\tau + d\tau$ равно числу объектов, рожденных за такой же интервал времени τ лет назад, но не всех, а только тех, для которых время жизни $T \geqslant \tau$. Таким образом,

$$g(\tau) = n(-\tau) \int_{\tau}^{\infty} f_0(T)\, dT\,, \tag{11.21}$$

что является условием выполнимости рассуждений, приведенных выше для произвольного $d\tau$.

Полученное интегральное уравнение (11.21) содержит те же две неизвестные функции, что и (11.15), так что, объединив эти два уравнения в систему, можно найти скорость рождения $n(t)$ и начальное распределение по временам жизни $f_0(T)$ для изучаемых объектов. Пусть теперь снова $n(t) = n_0 = \text{const}$. Уравнение (11.21) при этом переходит в уравнение

$$g(\tau) = n_0 \int_{\tau}^{\infty} f_0(T)\, dT\,, \tag{11.22}$$

то есть в интегральное уравнение, содержащее только одну неизвестную функцию $f_0(T)$ и один неизвестный параметр n_0. Дифференцируя это выражение по τ и заменяя затем переменную τ величиной той же размерности T, найдем решение

$$f_0(T) = -\frac{1}{n_0}\, \frac{d\,(g(T))}{dT}\,. \tag{11.23}$$

При выводе формул этого параграфа не принималось во внимание, что самые долгоживущие объекты Галактики могут иметь времена жизни, превосходящие возраст самой Галактики τ_g. Учет этого обстоятельства не изменит выражений (11.21)—(11.23). Формулы (11.15)—(11.20) тоже останутся справедливыми, но только для объектов с временем жизни $T < \tau_g$, для остальных же объектов все величины T, не яв-

ляющиеся аргументами функций $f_0(T)$ и $f(T)$, следует заменить на τ_g. А выражения (11.18) и (11.19) для $T > \tau_g$ не имеют места. Таким образом, для класса объектов, часть из которых имеет времена жизни больше возраста Галактики, при анализе их начального распределения по времени жизни приходится использовать различные формулы раздельно для объектов с $T < \tau_g$ и $T > \tau_g$.

12. МЕТОД НАИБОЛЬШЕГО ПРАВДОПОДОБИЯ И ЕГО ПРИМЕНЕНИЕ В ЗАДАЧАХ ЗВЕЗДНОЙ АСТРОНОМИИ

12.1. Метод наибольшего правдоподобия

Перейдем к важному для приложений методу получения оценок параметров выборки — **методу наибольшего правдоподобия**. Этот метод, основанный на подходах, предложенных Т. Байесом [35], активно развивал и пропагандировал Р. Фишер [36]. Метод основывается на следующем положении: ищутся такие оценки, которые максимизируют величину **правдоподобия** L некоторой выборки.

Определение 12.1. *Правдоподобием L выборки, содержащей n реализаций случайной величины X: $x_1,\ x_2,\ \ldots,\ x_n$, называется совместная плотность вероятности распределения $f(x_1,\ x_2,\ \ldots,\ x_n)$.*

Пусть L есть функция некоторого числа неизвестных параметров $\Theta_1,\ \ldots,\ \Theta_k$. Тогда оценками максимального правдоподобия этих параметров называются оценки $\breve{\Theta}_1,\ \ldots,\ \breve{\Theta}_k$, которые максимизируют величину L.

В случае независимых $x_1,\ x_2,\ \ldots,\ x_n$ совместная плотность распределения вероятностей есть произведение индивидуальных плотностей, таким образом получаем

$$L = \prod_{i=1}^{n} f(x_i)\,. \tag{12.1}$$

Рассмотрим пример, когда случайная величина X распределена по нормальному закону. В этом случае для i-й реализации данной случайной величины плотность распределения вероятности есть нормальное распределение

$$f(x_i) = K \exp\left(\frac{(x_i - \langle x\rangle)^2}{2\sigma^2}\right). \tag{12.2}$$

Функция правдоподобия при этом имеет вид

$$L = \prod_{i=1}^{N} f(x_i) = K^N \prod_{i=1}^{N} \exp\left(\frac{(x_i - \langle x\rangle)^2}{2\sigma^2}\right). \tag{12.3}$$

Перейдем к более удобному на практике натуральному логарифму функции правдоподобия, опустив при этом постоянный (в случае постоянной дисперсии) коэффициент, который нам далее не понадобится, при этом получим

$$\ln L = \sum_{i=1}^{N} \frac{(x_i - \langle x\rangle)^2}{2\sigma^2}\,. \tag{12.4}$$

Потребуем, чтобы наши оценки доставляли максимум этой функции. В общем случае математически сложной функции поиск экстремума может быть осуществлен численными метода-

ми минимизации. Для поиска экстремума приравняем частные производные от логарифма функции правдоподобия по искомым параметрам нулю (ограничимся выводом только для оценки математического ожидания, оценка для дисперсии получается аналогично, проделайте выкладки сами):

$$\frac{\partial \ln L}{\partial \langle x \rangle} = -2 \sum_{i=1}^{N} \frac{(x_i - \langle x \rangle)^2}{2\sigma^2} = 0\,. \tag{12.5}$$

Отсюда видно, что оценкой математического ожидания для нормального распределения служит арифметическое среднее

$$\langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i\,. \tag{12.6}$$

12.2. Выделение членов скоплений по собственным движениям и лучевым скоростям

Рассеянные звездные скопления являются ключевым объектом для проверки теории звездной эволюции, так как представляют собой группы звезд разных масс и приблизительно одинаковых возрастов. Это означает, что диаграммы показатель цвета — светимость рассеянных звездных скоплений практически представляют собой изохроны, которые можно сравнивать с теоретическими изохронами, построенными по эволюционным трекам звезд разных масс. Однако для построения диаграмм показатель цвета — светимость по данным фотометрическим наблюдениям необходимо отделить члены скоплений

от звезд ближнего и дальнего фона, что особенно актуально для молодых скоплений, видимых обычно на богатом звездном фоне Галактики. Выделить члены скоплений можно с помощью кинематических критериев по собственным движениям и лучевым скоростям звезд, учитывая тот факт, что скопление как целое движется относительно окружающих его звезд.

Метод Сандерса

Рассмотрим выделение членов скопления по собственным движениям. При этом используют различные методы, мы же остановимся на наиболее употребительном в настоящее время *методе Сандерса*. В этом методе двумерная плотность распределения собственных движений звезд в площадке, содержащей скопление, представляется в виде суммы двух нормальных распределений, причем для представления движений звезд скопления используется сферически симметричная гауссова функция, а для звезд фона — эллиптическая с осями, параллельными осям μ_α и μ_δ (далее, для уменьшения числа индексов, обозначим эти величины буквами μ и v):

$$
\begin{aligned}
& f(\mu_i,\ v_i) = f^f + f^c = \\
& \quad = \frac{N_f}{2\pi\,\Sigma_\alpha\Sigma_\delta} \exp\left(-\frac{1}{2}\left(\frac{(\mu_i - X_f)^2}{\Sigma_\alpha^2} + \frac{(v_i - Y_f)^2}{\Sigma_\delta^2}\right)\right) + \\
& \quad + \frac{N_c}{2\pi\sigma_c^2} \exp\left(-\frac{1}{2}\,\frac{\mu_i^2 + v_i^2}{\sigma_c^2}\right),
\end{aligned}
\tag{12.7}
$$

где за начало координат принято движение центра скопления. Здесь N_f — число звезд фона в рассматриваемой площадке; N_c — число звезд скопления и $N = N_f + N_c$ — общее число

звезд в площадке; Σ_α и Σ_δ — дисперсии движений звезд фона; σ_c — дисперсия движений звезд скопления; X_f и Y_f — средние движения по осям координат для звезд фона. Для сокращения записи обозначим экспоненты, входящие в (12.7), буквами A и B. Метод наибольшего правдоподобия дает для распределения (12.7) следующие шесть уравнений, с помощью которых можно определить неизвестные параметры:

$$
\begin{aligned}
&1.\quad \sum_{i=1}^{N} \frac{1}{f}\left(\frac{A}{\Sigma_\alpha \Sigma_\delta} - \frac{B}{\sigma_c^2}\right) = 0\,;\\
&2.\quad \sum_{i=1}^{N} \frac{A}{f}\left(\mu_i - X_f\right) = 0\,;\\
&3.\quad \sum_{i=1}^{N} \frac{A}{f}\left(v_i - Y_f\right) = 0\,;\\
&4.\quad \sum_{i=1}^{N} \frac{A}{f}\left(\frac{\left(\mu_i - X_f\right)^2}{\Sigma_\alpha^2} - 1\right) = 0\,;\\
&5.\quad \sum_{i=1}^{N} \frac{A}{f}\left(\frac{\left(v_i - Y_f\right)^2}{\Sigma_\delta^2} - 1\right) = 0\,;\\
&6.\quad \sum_{i=1}^{N} \frac{B}{f}\left(\frac{\mu_i^2 + v_i^2}{\sigma_c^2} - 2\right) = 0\,.
\end{aligned}
\tag{12.8}
$$

Уравнения решают методом последовательных приближений. Задавая некоторые начальные значения параметров, сначала решают первое уравнение относительно одного из параметров, затем, используя это значение параметра, решают второе уравнение относительно следующего параметра и т. д. Как показала практика, процесс сходится уже после трех-пяти приближений.

После определения параметров распределения (12.7) можно оценить вероятность принадлежности звезды скоплению по следующей формуле:

$$p_i^c = \frac{f_i^c}{f_i^c + f_i^f} \,. \tag{12.9}$$

В изложенном методе Сандерса сделаны сильные предположения о равенстве нулю ковариаций плотности распределения собственных движений звезд фона и адекватности представления плотности распределения нормальным распределением, поэтому определяемые по формуле (12.9) вероятности членства следует рассматривать лишь как величину, близкую к этой вероятности, и использовать только для разделения выборки на две совокупности — звезд скопления и звезд фона. Это разделение проводят с помощью построения гистограммы частотного распределения оценок p_i^c и выбора граничной величины по форме этого распределения.

Задачу о выделении членов скопления по собственным движениям можно решить непосредственно методом нелинейной регрессии, построив двумерную гистограмму для распределения собственных движений звезд в площадке и определяя параметры распределения с помощью минимизации суммы квадратов отклонений выражения (12.7) от соответствующих значений гистограммы. Этот метод несколько более нагляден, в нем проще контролировать сходимость процесса к «разумной» оценке компонент вектора параметров распределения. Необходимо помнить, что такое решение зависит от варианта разбиения области, в которой строится гистограмма.

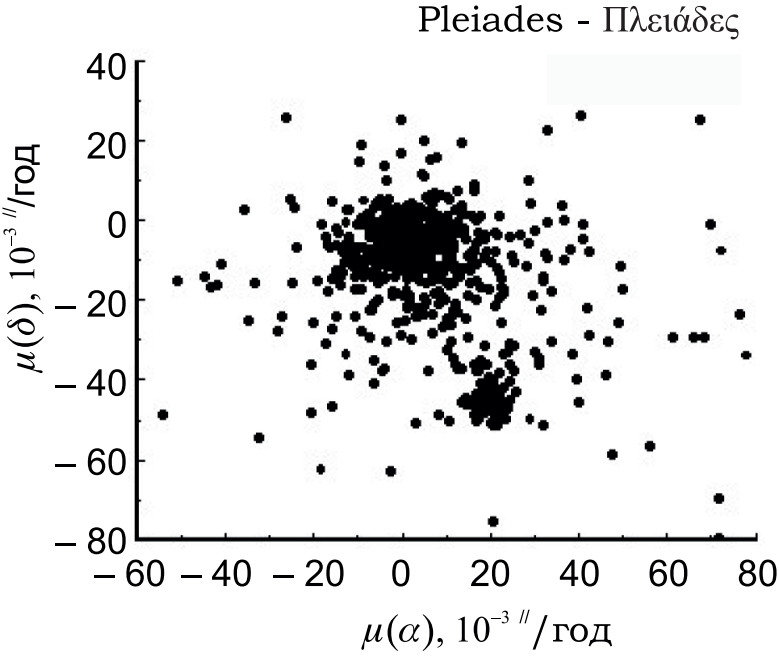

Рис. 12.1. Распределение собственных движений в области звездного скопления Плеяды

На рис. 12.1 показано распределение собственных движений звезд в области размером 3°, содержащей рассеянное звездное скопление Плеяды. Собственные движения для данной области неба взяты из каталога Tycho-2. На рисунке видно, что для этого близкого скопления выделение членов по собственным движениям можно провести достаточно уверенно: члены скопления группируются в ясно выделяющейся области в нижней части распределения.

Использование лучевых скоростей для выделения членов скопления в вычислительном смысле проще, так как для этой задачи используются одномерные распределения. Однако лишь для нескольких скоплений измерено достаточно большое количество лучевых скоростей, чтобы использовать решение, подобное изложенному выше, для собственных движений. Поэтому используют более простой подход: оценивают среднюю лучевую скорость скопления в целом, а затем для каждой звезды проверяют гипотезу об отличии лучевой скорости данной звезды от этого среднего.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. *Василевский А. Е.* Методы звездной статистики : учеб. пособие. Свердловск : Урал. гос. ун-т, 1985. 84 с.

   Пособие, послужившее основой для написания данного пособия.

2. *Курт Р.* Введение в звездную статистику. М. : Мир, 1969. 222 с.

3. *Schwarzschild K.* Ueber die Eigenbewegungen der Fixsterne // Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen. 1907. P. 614—632.

4. *Крамер Г.* Математические методы статистики. М. : Мир, 1975. 648 с.

   «Математические методы статистики» — классическое руководство по этой дисциплине. Впервые на русском языке оно было издано в 1948 г. и сыграло большую роль в развитии теоретических работ по математической статистике, а также в повышении уровня прикладных работ. Собственно математической статистике посвящена третья (последняя) часть книги, а ее вторая часть до сих пор является одним из лучших учебных пособий по теории вероятностей.

5. Статистика : учеб. для вузов / ред. И. И. Елисеева. СПб. : Питер, 2016. 368 с.

6. *Себер Дж.* Линейный регрессионный анализ. М. : Мир, 1980. 456 с.

7. *Шторм Р.* Теория вероятностей. Математическая статистика. Статистический контроль качества. М. : Мир, 1970. 368 с.

8. *Вентцель Е. С.* Теория вероятности : учеб. для вузов. М. : Высш. шк., 1999. 576 с.

9. *Болч Б., Хуань К.* Многомерные статистические методы для экономики. М. : Статистика, 1979. 317 с.

   Книгу стоит читать тем, кто хочет разобраться в теоретических основах методов многомерной статистики.

10. *Wishart J.* The generalized product moment distribution in samples from a normal multivariate population // Biometrika. 1928. Vol. 20. P. 32—52.

11. *Андерсон Т.* Введение в многомерный статистический анализ. М. : Гос. изд-во физ.-мат. лит., 1963. 500 с.

12. *Мардиа К.* Статистический анализ угловых наблюдений. М. : Наука, 1978. 240 с.

    «Статистический анализ угловых наблюдений» является по сути единственной монографией, где последовательно изложены вопросы статистической обработки величин, распределенных на окружности.

13. *Бабак В. П., Бабак С. В., Еременко В. С. и др.* Теоретические основы информационно-измерительных систем : учеб. / ред. В. П. Бабак. Киев : Издат. центр ТОВ «София-А», 2014. 832 с.

14. *Feigelson E. D., Babu G. J.* Modern Statistical Methods for Astronomy. Cambrige University Press, 2012. 476 p.

15. *Martinez V. J., Saar E.* Statistics of the Galaxy Distribution. Chapman, 2002. 430 p.

16. *Митропольский А. К.* Техника статистических вычислений. М. : Наука, 1971. 576 с.

    Несмотря на нетрадиционную терминологию, эта книга является ценным пособием по оцениванию формы и параметров наблюдаемых распределений и критериям значимости. В книге имеется обширное собрание статистических таблиц.

17. *Sturges H.* The choice of a class-interval // J. Amer. Statist. Assoc. 1926. Vol. 21. P. 65—66.

18. *Мостовова Н. А., Данилина Л. С.* Статистическая обработка экспериментальных данных при сертификации продукции. М. : Изд-во Рос. хим.-технол. ун-та им. Д. И. Менделеева, 2004. 21 с.

19. *Cooley J. W., Tukey J. W.* An algorithm for the machine calculation of complex Fourier series // Math. Comput. 1965. Vol. 19. P. 297—301.

20. *Гадзиковский В. И.* Цифровая обработка сигналов. М. : СОЛОН-Пресс, 2013. 766 с.

21. *Scheffe H.* A method for judging all contrasts in the analysis of variance // Biometrika. 1953. Vol. 40, iss. 1—2. P. 87—110.

22. *Хубаев Г. Н.* О влиянии ошибок независимых переменных на выбор состава факторов и структуры уравнения регрессии // Управление экономическими системами : электрон. науч. журн. 2010. Т. 3, вып. 23.

23. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ : в 2 т. М. : Финансы и статистика, 1986. Т. 1. 366 с. ; Т. 2. 251 с.

Этот небольшой двухтомник — лучшее пособие по классическому линейному регрессионному анализу из известных авторам. Здесь очень подробно и просто рассмотрены модели с одним и двумя предикторами, матричный подход также изложен просто, так что будет понятен и читателям, слабо знакомым с матричной алгеброй. В книге кратко изложены вопросы выбора наилучшего уравнения регрессии и регрессии в случае нарушения нормальности распределения ошибок в отклике. Издание также полезно приведенными в ней численными примерами, удобными для использования при отладке программ.

24. *Вучков И., Бояджиева Л., Солаков Е.* Прикладной линейный регрессионный анализ. М. : Финансы и статистика, 1987. 240 с.

Несколько сложная для понимания не очень опытными читателями из-за сжатости изложения, но очень полезная монография, в которой подробное рассмотрение стандартного линейного регрессионного анализа дополняется большим материалом по регрессионному анализу при плохо обусловленной информационной матрице, при коррелированных факторах и наличии ошибок в независимых переменных.

25. *Бард Й.* Нелинейное оценивание параметров. М. : Статистика, 1979. 349 с.

В довольно сложно написанной небольшой монографии подробно рассмотрены вопросы поиска экстремумов в нелинейных задачах.

26. *Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P.* Numerical recipes in C++ : the art of scientific computing. Cambridge University Press, 2002. 1235 p.

27. *Котельников В. А.* О пропускной способности эфира и проволоки в электросвязи // Материалы к I Всесоюз. съезду по вопросам техн. реконструкции дела связи и развития слаботоч. пром-сти : По радиосекции / Всесоюз. энергет. комитет. М. : Упр. связи РККА, 1933. С. 1—19.

28. *Shannon C. E.* Communication in the presence of noise // Proc. Institute of Radio Engineers. 1949. Vol. 37. P. 10—21.

29. *Nyquist H.* Certain topics in telegraph transmission theory // Trans. AIEE. 1928. Vol. 47. P. 617—644.

30. *Хемминг Р. В.* Цифровые фильтры. М. : Недра, 1987. 221 с.

    В монографии доступно и подробно рассмотрены принципы цифровой фильтрации и сглаживания данных. Даны методики расчета и анализа цифровых фильтров разных типов.

31. *Витязев В. В.* Вейвлет-анализ временных рядов : учеб. пособие. СПб. : С.-Петербург. гос. ун-т, 2001. 58 с.

    Пособие для студентов, содержит основы применения вейвлет-анализа для исследования случайных процессов.

32. *Кнут Д. Э.* Искусство программирования. Т. 2. Получисленные алгоритмы. М. : Вильямс, 2017. 832 с.

33. *Форсайт Дж., Малькольм М., Моулер К.* Машинные методы математических вычислений. М. : Мир, 1980. 280 с.

34. *Loktin A. V., Popova M. E.* Rotation curve of the Galaxy from the motions of open star clusters // Astronomical and Astrophysical Transactions. 2012. Vol. 27. P. 379—388.

35. *Bayes T.* An Essay towards solving a Problem in the Doctrine of Chances // Philosophical Transactions of the Royal Society of London. 1763. Vol. 53. P. 370—418.

36. *Fisher R. A.* On the mathematical foundations of theoretical statistics // Philosophical Transactions of the Royal Society of London. 1922. Vol. 222. P. 309—368.

37. *Фаддеев М. А.* Элементарная обработка результатов эксперимента : учеб. пособие. Нижний Новгород : Изд-во Нижегород. гос. ун-та, 2002. 108 с.

Приложение 1

Численные примеры к главе 1

Задача 1. Рассмотрим пример трехмерного случайного вектора $\vec{\mathbf{\Phi}}$, состоящего из трех компонент: фотометрических величин V, $B{-}V$ и $U{-}B$ для 15 звезд молодого рассеянного скопления Tr16 — наиболее ярких из полного массива, содержащего данные для 95 звезд. Каждая компонента случайного вектора представлена пятнадцатью ее реализациями, так что мы имеем матрицу данных размером 3×15, представляющую реализации нашего трехмерного случайного вектора фотометрических данных (табл. 1).

Таблицы 2 и 3 содержат ковариационную и корреляционную матрицы для таблицы данных 1. Напомним, что корреляционная матрица состоит из парных коэффициентов корреляции компонент случайного вектора $\vec{\mathbf{\Phi}}$, так что в нашем случае мы видим, что компонента V не связана линейно с компонентами $B{-}V$ и $U{-}B$, которые, в свою очередь, показывают близкую к линейной связь между собой.

Таблица 1

Матрица данных

V	$B-V$	$U-B$
7.37	0.16	-0.85
7.75	0.05	-0.91
7.82	0.17	-0.77
7.82	0.10	-0.70
8.10	0.41	-0.65
8.17	0.12	-0.84
8.42	0.10	-0.89
8.61	0.21	-0.78
8.77	0.14	-0.79
9.05	0.13	-0.86
9.29	0.32	-0.72
9.31	0.23	-0.77
9.31	0.31	-0.69
9.47	0.29	-0.75
9.53	0.10	-0.80

Таблица 2

Ковариационная матрица

0.518	0.019	0.013
0.019	0.010	0.007
0.013	0.007	0.006

Таблица 3

Корреляционная матрица

1.000	0.263	0.247
0.263	1.000	0.888
0.247	0.888	1.000

Задача 2. В следующем примере использован тот же источник данных для скопления Tr16, но взято уже 95 звезд с видимыми звездными величинами до $V = 14.0^m$. Получены следующие значения для ковариационной и корреляционной матриц (табл. 4, 5):

Таблица 4

Ковариационная матрица

2.604	0.100	0.386
0.100	0.017	0.024
0.386	0.024	0.092

Таблица 5

Корреляционная матрица

1.000	0.481	0.787
0.481	1.000	0.622
0.767	0.622	1.000

Сравните две корреляционные матрицы, помня что ковариации и коэффициенты корреляции подвержены влиянию случайных ошибок.

На рис. 1, 2 показаны для использованных данных диаграмма показатель цвета — звездная величина и двухцветная диаграмма, на которых видно, почему различаются компоненты корреляционной матрицы при переходе к большему массиву наших данных. Здесь светлыми кружками обозначены точки для первого массива, черными — для остальных данных.

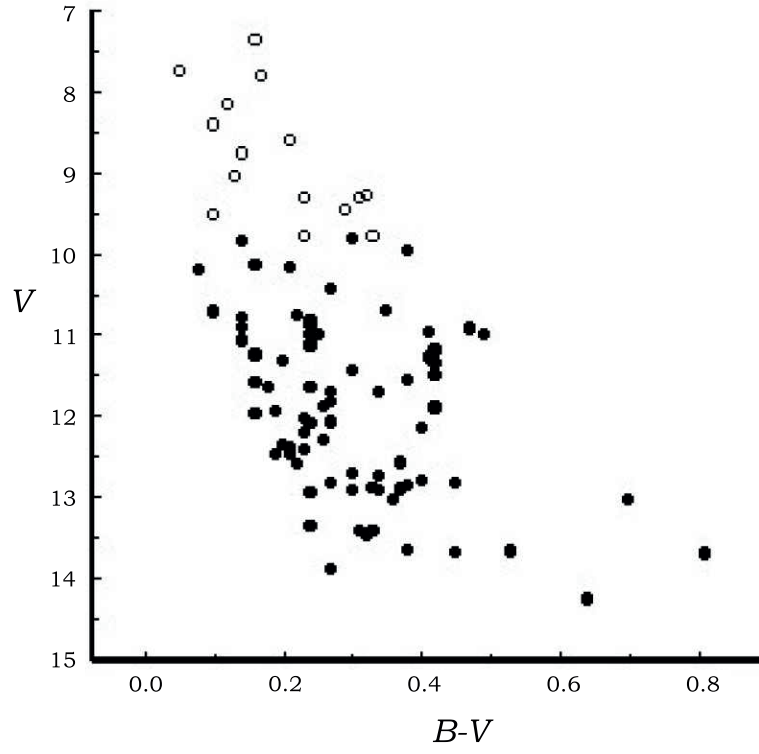

Рис. 1. Диаграмма показатель цвета — звездная величина

Задача 3. Мы приведем здесь пример вычисления множественного коэффициента корреляции величин V с величинами $B—V$ и $U—B$, хотя для вычисления его необходимо понимание регрессионного анализа. Поэтому к данному примеру желательно вернуться после изучения соответствующей главы. Итак, для первых пятнадцати строк данных (см. табл. 1) методом наименьших квадратов мы получаем зависимость

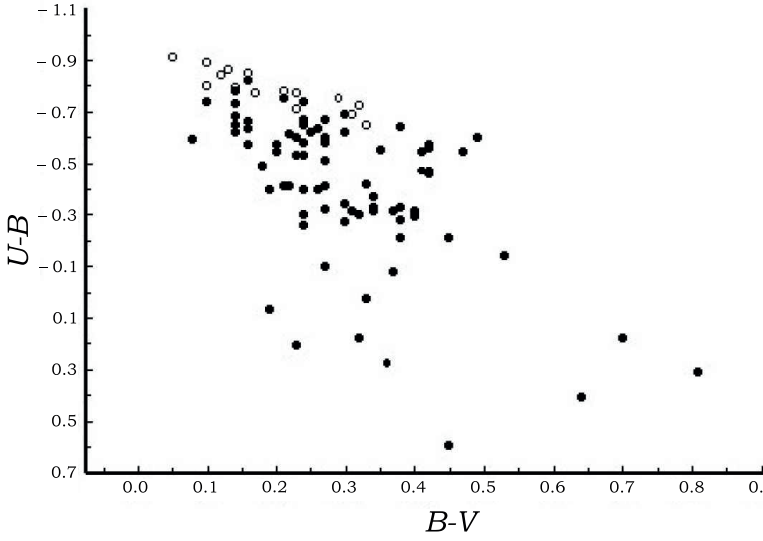

Рис. 2. Двухцветная диаграмма для звезд скопления

$$V = 8.75 + 1.51(B - V) + 0.58(U - B)\,. \qquad (1)$$

Общая дисперсия σ_V^2 величин V равна 7.25, сумма квадратов невязок σ^2 выражения (1), то есть сумма квадратов разностей левой и правой части (1), для всех строк табл. 1 равна 6.75. Затем по формуле

$$\rho^2_{V|BV|UB} = 1 - \frac{\sigma^2}{\sigma_V^2} \qquad (2)$$

получаем значение 0.07 для квадрата множественного коэффициента корреляции. Мы видим, что лишь незначительная часть общей дисперсии величин V объясняется линейной регрессион-

ной зависимостью с B—V и U—B, иными словами, величина V практически не коррелирует для данной выборки с показателями цвета.

Для более полной выборки (данные для всех 95 звезд) коэффициент детерминации $\rho^2_{V|BV|UB}$ уже равен 0.64, то есть уже около 80 % дисперсии в величине V объясняется регрессией — ее зависимостью от показателей цвета B—V и U—B.

Приложение 2

Простая линейная регрессия

Задача 1. В дополнение к рассуждениям о выборе МНК как наиболее удобном методе получения оценок регрессионных моделей, по крайней мере в случае линейных моделей, рассмотрим, чем отличаются методы получения оценок. На рис. 1 изображено множество точек, являющееся выборкой для получения коэффициентов модели, изображенной прямой линией. Одна из точек, как видно из рисунка, дальше других отходит от идеальной прямой. Как отклонение одной точки будет влиять на оценки, получаемые методом минимакса? Так как при применении метода минимакса минимизируется максимальное отклонение, то именно эта точка «подтянет» получаемую прямую к себе так, что отклонение этой точки сравняется с отклонениями нескольких точек, лежащих ниже кривой. Понятно, что принцип минимакса, где положение одной случайной точки критически влияет на оценку параметров прямой, совершенно не подходит для регрессионного анализа данных, отягощенных случайными ошибками.

В случае применения МНК одна отклоняющаяся точка тоже будет давать заметный вклад в минимизируемую сумму квадратов отклонений, но этот вклад не будет столь катастрофически влиять на результат оценивания параметров прямой. Эту точку впоследствии, во время обязательно заключающего процедуру оценивания анализа невязок, можно выделить как грубый промах и удалить из выборки. К сожалению, такой анализ невязок трудоемок при большом количестве параметров

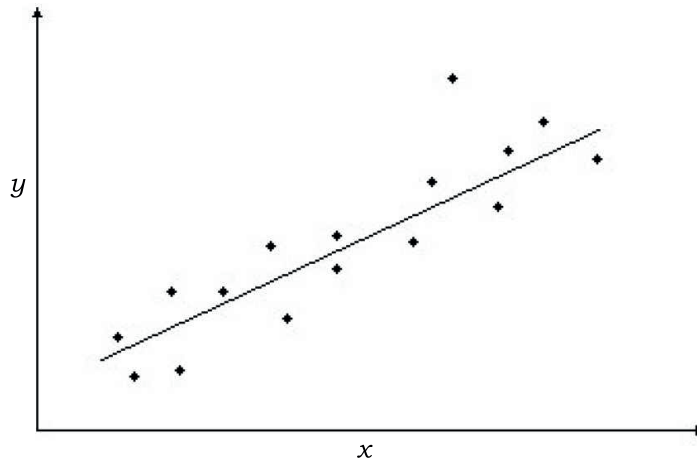

Рис. 1. Пример построения регрессионной зависимости для множества точек выборки

в модели. Ясно также, что МНК плох в случае асимметричного распределения невязок. В этом случае может помочь замена МНК на минимизацию суммы модулей меньших степеней уклонений. При таком подходе вклад отдельных далеко отстоящих точек еще уменьшается. Но минимизация сумм уклонений в степенях, отличных от второй, сильно усложняет решение регрессионной задачи. В большинстве случаев тщательный анализ распределения остаточных отклонений и приписывание отклоняющимся точкам меньших весов позволяет использовать МНК и в случае существенных отклонений распределения остаточных отклонений от нормального.

Задача 2. В качестве примера применения МНК рассмотрим однофакторное уравнение регрессии, то есть определение коэффициентов регрессионной модели, представленной прямой линией. Условные уравнения, записанные для n наблюденных точек, в этом случае выглядят следующим образом:

$$y_i = ax_i + b + \epsilon_i \,. \tag{3}$$

Чтобы получить систему нормальных уравнений, запишем сумму квадратов уклонений для уравнений (3):

$$SS = \sum_{i=1}^{n} (y_i - ax_i - b - \epsilon_i)^2 \,. \tag{4}$$

Эта функция от искомых параметров $(a,\ b)$ имеет минимум при значениях, когда частные производные от SS по a и b равны нулю:

$$\begin{aligned} \frac{\partial SS}{\partial a} &= -\sum_{i=1}^{n} x_i \,(y_i - ax_i - b - \epsilon_i) = 0 \,, \\ \frac{\partial SS}{\partial b} &= -\sum_{i=1}^{n} (y_i - ax_i - b - \epsilon_i) = 0 \,. \end{aligned} \tag{5}$$

Получается система нормальных уравнений — запись выражения (6.8) для случая двух факторов. Выражения (5) перепи-

шем в виде

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0\,, \\ \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - bn = 0\,. \end{aligned} \tag{6}$$

Меняя знаки и перенося члены, содержащие y_i, в правые части уравнений, получаем стандартный вид нормальных уравнений, которые и надо решить для получения оценок параметров a и b.

$$\begin{aligned} a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i\,, \\ a \sum_{i=1}^{n} x_i + bn = a \sum_{i=1}^{n} y_i\,. \end{aligned} \tag{7}$$

Можно легко решить эту систему обычным способом, но так как нам впоследствие все равно потребуется вычислять обратную матрицу $\left(\hat{\mathrm{X}}^T \hat{\mathrm{X}}\right)^{-1}$, найдем ее и получим оценки параметров с ее помощью. Выпишем лишь определитель матрицы системы нормальных уравнений:

$$\det = n \sum_{i=1}^{n} x_i^2 - \left(a \sum_{i=1}^{n} x_i \right)^2 , \tag{8}$$

именно значение этой величины определяет хорошую или плохую обусловленность матрицы условных уравнений, а значит, и решения всей задачи.

Итак, определим элементы матрицы, обратной к матрице системы условных уравнений. Для этого представим очевидное свойство обратной матрицы по компонентам:

$$\begin{pmatrix} \sum\limits_{i=1}^{n} x_i^2 & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} z_1 & z_2 \\ z_3 & z_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} . \tag{9}$$

Здесь первая матрица есть матрица системы нормальных уравнений, вторая — неизвестная обратная матрица, а их произведение есть единичная матрица согласно определению обратной матрицы. Перемножая матрицы в левой части (9), мы можем получить систему уравнений для определения величин z_i. Но еще лучше в данном случае просто выписать элементы обратной матрицы, пользуясь определением: элемент обратной матрицы есть алгебраическое дополнение элемента матрицы нормальных уравнений, стоящего на данном месте, деленное на определитель этой матрицы, так что

$$\begin{aligned} z_1 &= \frac{n}{\det} , \\ z_2 &= -\frac{\sum\limits_{i=1}^{n} x_i}{\det} , \\ z_3 &= -\frac{\sum\limits_{i=1}^{n} x_i}{\det} , \\ z_4 &= \frac{\sum\limits_{i=1}^{n} x_i^2}{\det} . \end{aligned} \tag{10}$$

Как видим, обратная матрица получилась симметричной, а на главной диагонали стоят строго положительные величины. Здесь величина det определяется выражением (8). Теперь для определения коэффициентов регрессионной модели используем выражение (6.9), но сначала запишем произведение матриц $\hat{\mathrm{X}}^T\hat{\mathrm{Y}}$. Это, очевидно, вектор, имеющий две компоненты. Так как матрица $\hat{\mathrm{X}}$ содержит два столбца, первый из которых содержит единицы, а второй x_i, то в результате умножения мы имеем компоненты $\sum_{i=1}^{n} x_i y_i$ и $\sum_{i=1}^{n} y_i$. Умножая обратную матрицу на этот вектор, получаем компоненты вектора оценок коэффициентов регрессионной модели:

$$a = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{\det}\,, \tag{11}$$

$$b = \frac{\sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i}{\det}\,. \tag{12}$$

Построение таблицы дисперсионного анализа и оценивание погрешностей коэффициентов уравнения регрессии и коридора погрешностей для предсказанных значений также не представляет трудности, и мы оставляем выкладки читателям в качестве упражнений. Процедура оценки погрешностей коэффициентов регрессии подробно представлена в [37], например.

ОГЛАВЛЕНИЕ

Учебное издание

Локтин Александр Васильевич
Островский Андрей Борисович

МЕТОДЫ ЗВЕЗДНОЙ СТАТИСТИКИ

Учебное пособие

Заведующий редакцией	М. А. Овечкина
Редактор	Т. А. Федорова
Корректор	Т. А. Федорова
Оригинал-макет	А. Б. Островский

Подписано в печать 10.04.2018 г. Формат $60\times84^{1}/_{16}$.
Бумага офсетная. Цифровая печать. Усл. печ. л. 14,64.
Уч.-изд. л. 8,5. Тираж 50 экз. Заказ 87.

Издательство Уральского университета.
Редакционно-издательский отдел ИПЦ УрФУ
620 083, Екатеринбург, ул. Тургенева, 4.
Тел.: +7 (343) 389-94-79, 350-43-28
E-mail: rio.marina.ovechkina@mail.ru

Отпечатано в Издательско-полиграфическом центре УрФУ
620 083, Екатеринбург, ул. Тургенева, 4.
Тел.: +7 (343) 358-93-06, 350-58-20, 350-90-13
Факс : +7 (343) 358-93-06
http://print.urfu.ru

УНИВЕРСИТЕТСКОГО КНИГОИЗДАНИЯ
УНИВЕРСИТЕТСКАЯ КНИГА
2015

9 785799 623159